Incorporating reliability measurements into the predictions of a recommender system

Antonio Hernando , Jesús Bobadilla, Fernando Ortega, Jorge Tejedor

Universidad Politécnica de Madrid, Filmaffinity Research Team, Spain

A B S T R A C T

In this paper we introduce the idea of using a reliability measure associated to the predictions made by recommender systems based on collaborative filtering. This reliability measure is based on the usual notion that the more reliable a prediction, the less liable to be wrong. Here we will define a general reliability measure suitable for any arbitrary recommender system. We will also show a method for obtaining specific reliability measures specially fitting the needs of different specific recommender systems.

1. Introduction

Recommender systems are programs that recommend to users a set of articles or services that they might interested them. Such programs have become popular due to the fast adoption of Web 2.0 [32] and the explosion of available information on the Internet. Although recommender systems cover a wide variety of possible applications [2,5,28,31], movie recommendation websites are perhaps the best-known example for common users; therefore, they have been subject to significant research [4,8].

Recommender systems are based on a filtering technique that attempts to reduce the amount of information available to the user. To date, collaborative filtering is the most commonly used and studied technology [1,9,15]; thus, a judgment on the quality of a recommender system depends significantly on its collaborative filtering procedures [15]. The different methods on which collaborative filtering is based are typically classified as follows:

- Memory-based methods [22,36] use similarity metrics to act directly on the matrix that contains the ratings of all users who have expressed their preferences on the collaborative service; these metrics mathematically express the distance between two users or two items based on their respective ratings.
- Model-based methods [1] use the user rating matrix to create a model on which the sets of similar users will be established. Among the most widely used model-based methods are the Bayesian classifiers [10], the genetic algorithms [7], the neural networks [17] and the fuzzy systems [37].

Generally, commercial recommender systems (e.g., Epinions, Netflix, FilmAffinity, LibimSeTi, Yahoo) use memory-based methods, and model-based methods are usually associated with research recommender systems. Regardless of the approach used in the collaborative filtering stage, the technical purpose generally pursued is to minimize the prediction errors, by

increasing the accuracy [3,14,33] of the recommender systems as high as possible. This accuracy is usually measured by the Mean Absolute Error (MAE) [1,6,17].

In this paper, we will focus on the memory-based methods that rely on the user-based nearest neighborhood algorithm [1,9]. The K users that are most similar to one given (active) user are selected on behalf of the coincidence ratio between their votes as registered in the database.

Our paper is concerned with a reliability measurement that is designed to improve the prediction information provided to users. Each prediction value relates to a reliability value that informs how 'likely' the prediction is to be correct. Accordingly, when recommending an item, we will provide the active user with two values: (1) the prediction about how much he will like this item and (2) the reliability (as we will see this is a number between 0 and 1) of this prediction. For example, a film recommender system, such as 'MovieLens' or 'NetFlix', could recommend to a user the film 'The Godfather' with the value (4.7, 0.8) which means that the recommender system predicts with a high reliability (a value of 0.8 over 1) that the user will like this film (with a rating of 4.7 over 5). Users could use these two values to make a balance between the prediction made by the recommender system and the reliability of the prediction.

Thus far, users in a recommender system (RS) are provided with recommendations and predictions about items based on a numerical value related to the prediction about how much the user would like the item (e.g., Titanic 4.2; Avatar 4.4; Australia 4.6). Consequently, a user in the RS chooses just the item with the highest numeric value; in the previous example, the user would choose Australia with a 4.6 value. However, this model does not offer a realistic representation of the way people recommend items to others. To consider a recommendation seriously, you take into account factors such as the number of people making the same recommendation, the similarity in taste these people share with you, and the agreement between these people. Real life users take into account the numeric value associated with the recommended item, as well as the manner in which this value has been obtained. Returning to the previous example, the following scenarios could occur:

- The recommendation for the film "Titanic" is based on the fact that there are 140 users with similar tastes to the active user who have rated the film highly.
- The recommendation for the film "Avatar" is based on the fact that there are only two users with similar tastes to the active user who have rated the film highly.

In this scenario, the user might be inclined to prefer "Titanic" over "Avatar" after becoming aware that the first recommendation was based on a wider number of users; therefore, the first recommendation could seem more reliable.

In this paper, we propose a reliability measure that forms a two-dimensional recommendation model. Recommendations within this model are established based on two numerical values: (1) a prediction of how much the user will like a given item and (2) a value representing the reliability of the prediction. In this manner, our reliability measure improves the one-dimensional model of present recommender systems that are based on collaborative filtering. Using our model, users may make their choice by considering these two values.

We believe that this reliability measure is important both for the users of recommender systems and for the researchers of this area:

Importance for users.
When evaluating predictions and recommendations, users initially consider all of them with the same degree of reliability. To understand the comparative reliability of a prediction, users must be acquainted with the K-nearest neighbors algorithm on which many recommender systems (based on collaborative filtering) are based and also have information about the neighbors and their ratings. Prediction reliability information provides this information directly to users.

Importance for researchers.
The reliability measure provides a method of evaluating the quality of predictions. According to this reliability measure we could compare the quality obtained by using different metrics.

The reliability measure proposed in this paper can be regarded as a memory-based quality measure because it is calculated by taking into account only the ratings that users make for the items; it is not calculated by taking into account any other type of information such as the trust of a user. In this way, the proposed reliability measure can be applied to any recommender system based on collaborative filtering because all of those systems are based on user ratings; however, not all of the recommender systems based that on collaborative filtering are provided with information about the trust between users, the demographic data, the content-based data, etc.

In Section 2 we discuss the main difference between our approach and other related works. In Section 3 we formalize some concepts on recommender systems based on collaborative filtering. In Section 4 we present a general scheme for defining a reliability measure that is associated with the predictions. Using this scheme, in Section 5, we define a reliability measure suitable for any recommender system that is based on collaborative filtering. In Section 6 we prove this reliability measure using two known recommender systems (*MovieLens* and *NetFlix*). In Section 7, we examine how the reliability measure may be used to compare the quality of different metrics. Finally, in Section 8, we draw our conclusions.

2. Related work

The concept of reliability measure associated with predictions is new in the field of collaborative filtering (in the manner it is applied in this paper). To date, there are two main research lines in the literature that are different from but closely related to ours:

Explaining recommendations and pervasive collaborative filtering.
A small number of papers belong to this line. The following studies are among the most outstanding: [33] presents new results in developing design principles and algorithms for constructing explanation interfaces. They provide a new interface called the organization interface in which the results are grouped according to their tradeoff properties. [30] proposes a case-based reasoning approach to product recommendation that offers benefits in the ease with which the recommendation process can be explained and how system recommendations can be justified. The goal of [35] is to examine the role of transparency (the user understanding of why a particular recommendation was made) in recommender systems. To explore this issue, the authors conducted a user study of five music recommender systems. In [14] the authors address the explanation interfaces for ACF systems (how and why they should be implemented).

Trust and reputation.
Many papers focus on terms with the same meaning such as trust, credibility, reliability and reputation (as summarized in Table 1). At present, the term 'trust' tends to be used most often. Nevertheless, our approach is different from all of these papers, which associate a measure with an item or a user. Instead, our approach associates a measure with a prediction made by the RS of how much a user would like an item.
The recommender systems obtain the 'trust' measure associated with users through explicit ratings provided by other users in [12,33,34]. The P2P services often use this method (see [25]). In [11,29], the 'trust' measure is calculated through the implicit relations of users in a social network.
In [20] the trust associated with items is calculated through a feedback of users who are asked their opinion of the items. This method is often used by e-Commerce services. In [11,21] the 'trust' associated with items is calculated through the implicit information provided by users (e.g., the number of times a user has listened to a song, or the number of times a user has visited a web page).
In [19,23] the 'trust' measure associated with items and users is calculated by taking into account only the user ratings. In the field of collaborative filtering, the trust of users is used to make predictions by weighting the trust of the users. The more trust a user has, the more important his ratings are for making predictions (see [11,18,33]).

Note the two common features of the works examining these 'trust' measures associated with users or items:

(a) The 'trust' measure is calculated by analyzing the information (either implicit or explicit) of the rest of users.
(b) The main purpose of this 'trust' measure is to improve the accuracy (MAE) of the recommender system using the 'trust' of users or items.

Conversely, the reliability measure associated with predictions in this paper has the two following features:

(a) The reliability measure associated with a prediction made to a user is calculated by taking into account only the information provided by its k-neighbors.
(b) The main purpose of the reliability measure is to improve the trust users have on predictions instead of improving the accuracy of the recommender system.

Fig. 1 shows a diagram on which we have placed the 'trust' (or reputation) measure and our reliability measure on behalf of these features. The 'trust' measure associated with users or items is placed in the "users relations level" or "to improve

Table 1
State of the art on trust and reputation.

	User trust	Item trust
Explicit trust systems	The 'credibility' of users is calculated by the explicit information of other users [12,33,34]. P2P services usually implement this technique [25]	The 'reputation' of items is calculated using the feedback of users who are asked about their opinions [20]. Services such as e-commerce often use this technique
Implicit trust systems	The 'credibility' of users is calculated using the implicit information obtained in a social network [11,29]	The 'reputation' of items is calculated by studying how users work with these items (e.g., the number of times a song is played) [11,21]
Memory based trust	The 'credibility' measure is calculated by taking into account the users' ratings [19,23]	

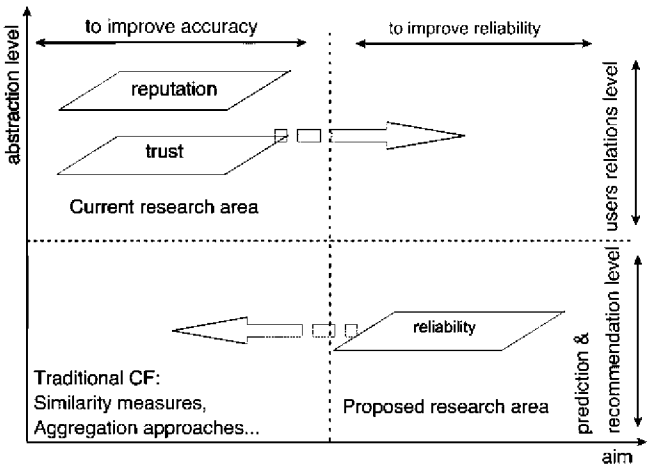

Fig. 1. The proposed reliability measure context.

accuracy", whereas our reliability measure associated with predictions is placed in the "predictions and recommendations level" and "to improve credibility". The horizontal arrows inform about the open research areas that could be examined in future works: the 'trust' measure could be used for improving prediction credibility; our reliability measure could be used for improving predictions and the accuracy of recommendations.

The proposed reliability measure may also be used in those RS including the additional information that is related to users (demographic information, trust and distrust relations, content-based, etc.). However, we emphasize that we could define a reliability measure based on this additional information for these RS.

3. Collaborative filtering based on the *k*-nearest neighbor algorithm

In this section we will describe the main concepts on which recommender systems and collaborative filtering are based.

We will consider a recommender system based on a database consisting of a set of m users, $U = \{1, \ldots, m\}$ and a set of n items, $I = \{1, \ldots, n\}$. In the case of a movie recommender system, U would stand for the database users registered in the system and I would refer to the different movies in the database.

Users rate those items they know with a discrete range of possible values $\{min, \ldots, max\}$ and associate higher values with their favorite items.[1] This range of values is typically $\{1, \ldots, 5\}$ or $\{1, \ldots, 10\}$.

Given a user $u \in U$ and an item $i \in I$, the expression $r_{u,i}$, will stand for the value with which the user u has rated the item i. Obviously, users may not have rated every item in I. We will use the symbol $\bullet$ to represent that a user has not made any rating concerning an item i. The set $\{min, \ldots, max\} \cup \{\bullet\}$ represents the possible values in the expression $r_{u,i}$.

To offer reliable suggestions, recommender systems try to accurately predict how a user u would rate an item, i, which has not yet been rated by the user u ($r_{u,i} = \bullet$). Given a user $u \in U$ and an item $i \in I$, we will use the expression $p_{u,i}$, to denote the system estimation of the value with which the user u is expected to rate the item i. The idea for calculating this estimation $p_{u,i}$ in collaborative filtering is based on the following: if we find a user $v \in U$ who has rated similarly to $u \in U$, then we can conclude that the tastes of the user u are akin to those of the user v. Consequently, given an item $i \in I$ which the user u has not yet rated although the user v already has, we could infer that the user u would most likely rate the item i with a similar value to the rating given by the user v.

Thus, in the collaborative filtering searches for each active user $u \in U$, a subset of k users,[2] $K_u = \{v_1, \ldots, v_k\} \subseteq U$ (called 'neighbors') who has rated similarly to the user u. To predict the value with which the user u would rate an item $i \in I$, the recommender system first examines the values with which the neighbors $v_1, \ldots, v_k$ have rated the item i, and then, uses these values to make the prediction $p_{u,i}$. Consequently, two main issues must be considered to make predictions:

- Evaluate how similar two users are to select, for each user u, a set of users $K_u = \{v_1, \ldots, v_k\} \subseteq U$ (called its 'neighbors') who have similar tastes to the user u.

[1] Although the users in most of the memory-based recommender systems rate with a range of quantitative values, there are situations in which the information cannot be assessed precisely in a quantitative form but may be assessed in a qualitative form. The use of the Fuzzy Sets Theory has provided good results for modeling qualitative information [32]. Fuzzy linguistic recommender systems [31] are based on the fuzzy linguistic modeling, allowing us to represent and handle the subjectivity, vagueness and imprecision that characteristically occur in processes of information searching and in this way the developed systems grant users the access to quality information in a flexible and user-adapted manner[16].

[2] In this approach it is difficult to decide the most suitable value k to be chosen. It usually depends on the metric and the recommender system. If we choose k as the number of users of the recommender system, we will make predictions based on the ratings of the whole set of users. However, the approach based on rating the whole collection of users provides worse results [34].

Table 2
The ratings made by users.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}
u_1	1	2	•	4	2	•	3	4	•	4	1	2	4	5	1
u_2	1	•	4	5	1	5	3	4	1	5	2	1	4	5	1
u_3	1	2	5	2	•	1	•	3	4	5	2	•	1	2	5
u_4	2	1	4	4	1	•	3	5	5	4	2	1	5	4	1
u_5	2	2	4	•	1	•	3	•	4	•	2	1	1	2	5
u_6	1	•	5	2	1	•	2	4	•	4	3	2	1	3	4
u_7	2	•	•	4	2	•	5	1	•	1	5	1	1	1	5
u_8	2	2	4	4	1	•	2	5	•	5	1	•	4	5	1
u_9	5	1	•	1	5	2	5	5	•	4	1	5	2	3	1

- Given a user u, and an item i, estimate the value, $p_{u,i}$, with which user u would rate item i by considering the values $r_{v_1,i},\ldots,r_{v_k,i}$ with which the neighbors of u have rated item i.

Regarding the first issue, there are several possible measures for quantifying the similarity of the ratings between two different users (through a function between users *sim*). Although different similarity measures have been proposed, the correlation coefficient or Pearson similarity is most commonly used. Once a similarity measure has been chosen, the recommender system selects a subset of the k users most similar to it, $K_u = \{v_1,\ldots,v_k\} \subseteq U$, for each active user u.

For the second issue, the most method of determining $p_{u,i}$ is the following aggregation approach (deviation from mean):

$$p_{u,i} = \bar{r}_u + \frac{\sum_{v\in K_{u,i}} sim(u,v)\cdot(r_{v,i}-\bar{r}_v)}{\sum_{v\in K_{u,i}} sim(u,v)} \tag{1}$$

where $K_{u,i} = \{v \in K_u | r_{v,i} \neq \bullet\}$ is the set of neighbors of u who have rated the item i; and $\bar{r}_u$ is the average rating made by the user u.

A way to measure the quality of these estimations is to calculate the mean absolute error [15] which conveys the mean of the absolute difference between the real values rated by the users, $r_{u,i}$, and the estimated values $p_{u,i}$:

$$MAE = \frac{\sum_{(u,i)\in J}|r_{u,i}-p_{u,i}|}{|J|}$$

where $|J|$ is the set of predictions $p_{u,i}$ that the recommender system can make and such that $r_{u,i} \neq \bullet$.

3.1. Example

Now we will consider a small example to clarify the concepts described above.

Consider a recommender system with nine users, $U = \{u_1,\ldots,u_9\}$ and 15 items, $I = \{i_1,\ldots,i_{15}\}$. Users rate items with the discrete set of values $\{1,\ldots,5\}$, as observed in Table 2. Thanks to this table, we can use a similarity measure to calculate the similarity between the users; in this case, we have used correlation. The resulting similarity between the users is shown in Table 3.

We will consider in this example that $K = 3$ and we would like to make a recommendation to user u_1. The 3 neighbors of user u_1 are $K_{u_1} = \{u_2, u_4, u_8\}$ because they are the three most similar users to u_1 (see Table 3). Thanks to these three neighbors, by means of the formula (1) the recommender system can make a prediction, $p_{u_1,i}$, of how much the user u_1 will like the items this user has not yet rated:

- The three neighbors of user u_1 have all rated the item i_3 ($K_{u_1,i_3} = \{u_2,u_4,u_8\}$),with value **4** (all of them find i_3 interesting). Consequently, the recommender system will predict that the user u_1 will like the item i_3 (with value **4**, $p_{u_1,i_3} = \mathbf{4}$).
- Only the neighbor u_2 of the user u_1 has rated the item i_6 ($K_{u_1,i_6} = \{u_2\}$), and this user has rated the item i_6 with value 5. In this case, the recommender system will predict that the user u_1 will like the item i_6 (with value 5, $p_{u_1,i_6} = 5$).
- Only the neighbors u_2 and u_4 of the user u_1 have rated the item i_9 ($K_{u_1,i_9} = \{u_2,u_4\}$), and they have rated it with the values 1 and 5, respectively. Consequently, the recommender system will predict that u_1 will like the item i_9 with the value:

$$p_{u_1,i_9} = 2.75 + \frac{0.92\cdot(1-3)+0.85\cdot(5-3)}{0.92+0.85} = 2.67$$

4. A framework for obtaining reliability measures

In this section we will discuss a general outline for defining a reliability measure $R_{u,i}$ for a prediction $p_{u,i}$. This section will be useful in the next section where we will present a specific reliability measure that is suitable for use in any recommender

Table 3
The similarity between users.

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
u_1	–	0.92	−0.02	0.85	−0.30	0.16	−0.49	0.91	0.09
u_2	0.92	–	−0.15	0.69	−0.07	0.35	−0.27	0.91	−0.27
u_3	−0.02	−0.15	–	0.06	0.96	0.89	0.17	0.11	0.04
u_4	0.85	0.69	0.06	–	0.17	0.26	−0.37	0.90	0.05
u_5	−0.30	−0.07	0.96	0.17	–	0.79	0.73	−0.11	−0.37
u_6	0.16	0.35	0.89	0.26	0.79	–	0.10	0.34	−0.22
u_7	−0.49	−0.27	0.17	−0.37	0.73	0.10	–	−0.69	−0.50
u_8	0.91	0.91	0.11	0.90	−0.11	0.34	−0.69	–	0.16
u_9	0.09	−0.27	0.04	0.05	−0.37	−0.22	−0.50	0.16	–

system based on collaborative filtering. We believe that the outline presented here could be used for defining a reliability measure fitted for a particular and specific recommender system.

We define a reliability measure as a real value between 0 and 1 that is associated with each prediction the recommender systems makes[3] in such a manner that the more reliability a prediction has, the more unlikely it is to be wrong.

The general outline described here is based on the idea of considering that the reliability, $R_{u,i}$, of any prediction, $p_{u,i}$ is calculated by taking into account different factors $F_{u,i}^{(1)} \ldots, F_{u,i}^{(N)}$. A factor $F_{u,i}$ is a real number that depends on the prediction $p_{u,i}$ and can be calculated from the ratings made by the neighbors of u.[4] We will distinguish between two types of factors:

Definition 1. We will distinguish these two types of factors:

(i) A factor $F_{u,i}^{(n)}$ is *positive* if the following holds true: the greater the value of $F_{u,i}^{(n)}$ for a prediction $p_{u,i}$ (considering the other factors for this prediction unchanged), the greater the value of the reliability of the prediction. In Section 5.1 we will see how the factor 'the number of neighbors of u who have rated the item i' may be regarded as a positive factor for measuring the reliability of the prediction $p_{u,i}$.

(ii) A factor $F_{u,i}^{(n)}$ is *negative* if the following holds true: the greater the value of $F_{u,i}^{(n)}$ for a prediction $p_{u,i}$ (considering the other factors for this prediction unchanged), the lesser the value of the reliability of the prediction. In Section 5.2 we will see how the factor 'the disagreement degree between the neighbors of u when rating the item i' may be regarded as a negative factor for measuring the reliability of the prediction $p_{u,i}$.

Generally, the following steps are needed to formally define the reliability measure $R_{u,i}$ of a prediction $p_{u,i}$.

Identifying Factors.
The initial step for defining the reliability measure is to identify the factors that play an important role in defining of the reliability measure. The consideration of a factor for measuring the reliability must be motivated by both psychological and mathematical reasons.

The consideration of a factor must be motivated by psychological reasons because each factor that is considered for defining reliability must seem reasonable enough.

Nevertheless, the inclusion of a factor must also be motivated by mathematical reasons. To include a factor $F_{u,i}^{(n)}$ for measuring the reliability, we propose to test the factor through the following experiment. Obtain a random sample of predictions in the recommender systems and study the following:

(i) Obtain a scatter plot where each point $(F_{u,i}^{(n)}, e_{u,i})$ represents information of $F_{u,i}^{(n)}$ (the value of the factor for the prediction $p_{u,i}$) and the error $e_{u,i} = |r_{u,i} - p_{u,i}|$ of the prediction $p_{u,i}$. We observe the following through this scatter plot:
- If $F_{u,i}^{(n)}$ is a *positive factor*, then we must observe that the greatest errors must appear when $F_{u,i}^{(n)}$ takes low values (e.g., the greatest errors appear when *few* neighbors of u have rated the item i).
- If $F_{u,i}^{(n)}$ is a *negative factor*, then we must observe that the greatest errors appear when $F_{u,i}^{(n)}$ takes high values (e.g., the greatest errors appear when there is a *high* degree of disagreement between the neighbors of u over rating the item i).

[3] Once the reliability measure is defined over the range [0, 1], it could easily be transformed so that it could take other values such as $\{1, \ldots, 5\}$, which correspond to the different values a user may use to rate an item.

[4] As we will see in next section, $|K_{u,i}|$, the number of neighbors of u who have rated the item i, is a factor that may be used to measure the reliability of a prediction.

(ii) We will distinguish two cases.

- $F_{u,i}^{(n)}$ is a *positive factor*. In this case, we will define $\text{MAE}_{F^{(n)}>v}$ as the mean absolute error of the predictions $p_{u,i}$ such that $F_{u,i}$ is *greater* than v. We can obtain a plot depicting the evolution of $\text{MAE}_{F^{(n)}>v}$ as related to v (i.e. $\text{MAE}_{F^{(n)}>1}$, $\text{MAE}_{F^{(n)}>2}$, etc.). Because $F_{u,i}^{(n)}$ is a positive factor, we must observe that $\text{MAE}_{F^{(n)}>v}$ is a decreasing function (e.g., the MAE falls when considering only the predictions calculated through a *high* number of neighbors of u who have rated the item i).
- $F_{u,i}^{(n)}$ is a *negative factor*. In this case, we will define the $\text{MAE}_{F^{(n)}<v}$ as the mean absolute error of the predictions $p_{u,i}$ such that $F_{u,i}$ is *lesser* than v. We can obtain a plot (see Fig. 6) that depicts the evolution of $\text{MAE}_{F^{(n)}<v}$ as related to v. Because $F_{u,i}^{(n)}$ is a negative factor, we must observe that $\text{MAE}_{F^{(n)}<v}$ is an increasing function (e.g., the MAE falls when considering only the predictions over which there is a *low* disagreement degree between the neighbors of u over rating the item i).

Defining the partial reliability of a factor.
Once a factor $F_{u,i}^{(n)}$ is considered for measuring the reliability, we define the *partial reliability* of this factor. The partial reliability of the factor $F_{u,i}^{(n)}$ is the reliability of the prediction $p_{u,i}$ solely considering factor $F_{u,i}^{(n)}$ for measuring the reliability.

Definition 2. The partial reliability of a factor $F_{u,i}^{(n)}$ is just a function $f_n : \mathbb{R} \to [0,1]$ that fulfills the following:

(i) If the factor $F_{u,i}^{(n)}$ is positive, then f_n must be an increasing function.
(ii) If the factor $F_{u,i}^{(n)}$ is negative, then f_n must be a decreasing function.
(iii) $f_n(\bar{w}) = 0.5$ where $F_{u,i}^{(n)} = \bar{w}$ must be a value close to the median of the values of F_i in the specific recommender system.

The partial reliability of a factor must fulfill the restrictions (i) and (ii) because of the factor definition (see Definition 1). The restriction (iii) is demanded so that $f_n(F_{u,i}^n)$ does not often reach low values; note that the reliability measure is above all a psychological measure. Otherwise, the users of the recommender system would always obtain low values in the reliability of their predictions, which may induce them to distrust the recommender system.

Defining the importance of the factors.
Once we have studied the partial reliability of a factor $F_{u,i}^{(n)}$, we will study its importance on the global reliability of the prediction. We consider that not all factors may have the same importance for measuring the reliability of a prediction. Consequently, we must define the importance, α_n, of the factor $F_{u,i}^{(n)}$ as a real value. The importance of a factor may either be a constant (we estimate that the importance of the factor is always the same for any prediction) or a variable which depends on the prediction.

Defining the Reliability measure.
Finally, the reliability $R_{u,i}$ measure of a prediction $p_{u,i}$ is defined in terms of the partial reliability of the identified factors. Indeed, the proposed reliability $R_{u,i}$ is the geometric mean of the partial reliability of the factors that are considered weighted by the importance of the factors.

Definition 3 (*Reliability Measure*). Let $F_{u,i}^{(1)}, \ldots, F_{u,i}^{(N)}$ be the factors considered for measuring the reliability. The reliability $R_{u,i}$ of a prediction $p_{u,i}$ is defined as follows:

$$R_{u,i} = (f_1(F_{u,i}^{(1)})^{\alpha_1} \cdot \ldots \cdot f_N(F_{u,i}^{(N)})^{\alpha_N})^{\frac{1}{\alpha_1+\ldots+\alpha_N}}$$

According to this definition, the following desired property holds, which is why we have chosen the geometric mean instead of the arithmetic mean: if the partial reliability of a factor $F_{u,i}^{(n)}$ is 0 (that is to say, $f_n(F_{u,i}^{(n)}) = 0$), then the reliability is $R_{u,i} = 0$.

4.1. Example

We will now consider an example to clarify the concepts described above. In this example, we will propose a naive reliability measure, which will illustrate how the reliability measure can be defined and how it is calculated over the predictions made by the recommender system.

In this example, we will consider two factors involved in the reliability measure associated with predictions.

Identifying the factors.
We will consider two basic factors for measuring the reliability of a prediction $p_{u,i}$:

- $|K_{u,i}|$: The number of neighbors of u who have rated the item i. This is a positive factor because the more neighbors who have rated an item, the more reliable the prediction based on the choice of those neighbors'.
- $V_{u,i}$: The variance of the ratings made by the neighbors of u over the item i. This factor measures the degree of disagreement between the neighbors of u when rating the item i. This is a negative factor because the more disagreement there is between the neighbors of u, the resulting prediction will be less reliable.

Defining the reliability measure of the factors

- We will consider the following function as the partial reliability of the positive factor $|K_{u,i}|$. As shown, it fulfills the properties of Definition 2[5]:

$$f_K(|K_{u,i}|) = \frac{|K_{u,i}|^2}{9}$$

- We will consider the following function as the partial reliability of the negative factor $V_{u,i}$. As shown, it fulfills the properties of Definition 2:

$$f_V(V0_{u,i}) = \frac{1}{1+V_{u,i}}$$

Defining the importance of the factors.
In this example, we will consider that both factors have the same importance ($\alpha_{|K_{u,i}|} = \alpha_{V_{u,i}} = 1$).

Defining the reliability measure.
According to Definition 3, the reliability measure would be:

$$R_{u,i} = \sqrt{f_K(|K_{u,i}|) f_V(V_{u,i})} = \frac{|K_{u,i}|}{3\sqrt{1+V_{u,i}}}$$

Now, we will see how this reliability measure would be calculated for some predictions made in the recommender system described in Example 3.1. As we have already established, the recommender system could make the following predictions related to user u_1:

The prediction $p_{u_1,i_3} = 4$.
The recommender system will predict that user u_1 will like the item i_3 (with a value of 4). In this prediction, we show that $|K_{u_1,i_3}| = 3$ (there are three neighbors who rate the item i_3) and $V_{u_1,i_3} = 0$ (all of these neighbors agree with the rating for the item i_3). According to the previous formula, the reliability associated with the prediction, p_{u_1,i_3} would be:

$$R_{u_1,i_3} = \frac{|K_{u_1,i_3}|}{3\sqrt{1+V_{u_1,i_3}}} = \frac{3}{3\sqrt{1+0}} = 1$$

As shown, this prediction is reliable because all of the neighbors of u_1 have rated the item i_3, and all of them like item i_3.

The prediction $p_{u_1,i_6} = 5$.
The recommender system will predict that user u_1 will like the item i_6 (with a value of 5). In this prediction, we show that $|K_{u_1,i_6}| = 1$ (only one neighbor has rated the item i_6) and $V_{u_1,i_6} = 0$ (because only one neighbor has rated the item i_6). According to the previous formula, the reliability associated with this prediction would be:

$$R_{u_1,i_6} = \frac{|K_{u_1,i_6}|}{3\sqrt{1+V_{u_1,i_6}}} = \frac{1}{3\sqrt{1+0}} = 0.33$$

Consequently, this prediction is unreliable because only one of the neighbors of u_1 has rated the item i_6 (factor $|K_{u_1,i_6}| = 1$). Consequently, the recommender system would prefer to recommend the item i_3 to user u_1 rather than the item i_6 because it is more reliable, although the value of prediction $p_{u_1,i_6} = 5$ is higher than that of $p_{u_1,i_3} = 4$.

The prediction $p_{u_1,i_9} = 2.67$
. The recommender system will predict that the user u_1 will like the item i_9 with a value of 2.67. In this prediction, we show $|K_{u_1,i_9}| = 2$ (two neighbors have rated item i_9) and $V_{u_1,i_9} = 4$ (these two neighbors disagree over the rating of the item i_9). According to the previous formula, the reliability associated with this prediction would be:

[5] Observe that we have considered in this example that $k = 3$, and consequently $|K_{u,i}| \leqslant k = 3$

$$R_{u_1,i_9} = \frac{|K_{u_1,i_9}|}{\sqrt{1+V_{u_1,i_9}}} = \frac{2}{3\sqrt{1+4}} = 0.30$$

Consequently, this prediction is also unreliable because there is a high disagreement between the neighbors of the user u_1 (according to factor V_{u_1,i_9}).

Here we have defined a simple and naive reliability measure to clarify the concepts related to the steps for defining and applying a reliability measure. However, this reliability measure is not good enough to be applied in real recommender systems. In the next section, we will study a more sophisticated and useful reliability measure, which provides good results in real recommender systems based on the k-neighbors algorithm (as shown in Section 6).

5. The reliability of a prediction

Previously, we have depicted a scheme for defining the reliability measures of predictions. Here we will use this scheme to define a reliability measure that is suitable for any recommender system based on collaborative filtering. This reliability measure is based on two factors discussed in Section 5.1 and Section 5.2. To justify that both factors must be considered in any recommender system that is based on collaborative filtering, we will study a random sample of 2000 predictions in two important and well known recommender system databases: *MovieLens* and *Netflix*. To make predictions, we will use correlation as the similarity measure and a constant k of 500 neighbors.

As discussed, these two factors and the reliability measure based on them will be defined in Section 5.3. As we will show in Section 6, this reliability measure may be useful for any recommender system that is based on collaborative filtering. We emphasize again that the outline described in the previous section makes it easy to include more specific factors that are fitted for a particular recommender system.

5.1. The factor $S_{u,i}$

In this section, we will consider an important factor ($S_{u,i}$) for measuring the reliability of a prediction $p_{u,i}$ (the prediction of a rating that a user u would make for an item i), which is closely related to $|K_{u,i}|$, the number of neighbors of u who have rated the item i.

Before formally defining this factor $S_{u,i}$, we will discuss how at first glance the factor $|K_{u,i}|$ might be considered to be an important factor for measuring the reliability of a prediction $p_{u,i}$.

As previously stated, we must distinguish between K, the number of neighbors of a user u (a constant defined in the recommender system), and $|K_{u,i}|$, the number of neighbors of u who have really rated the item i. Note that some of these neighbors may not rate the item i. Indeed, although K is a constant for any prediction $p_{u,i}$, the number $|K_{u,i}|$ ranges between different values. In Fig. 2, we can observe the distribution of the value $|K_{u,i}|$ for the random sample of predictions in the MovieLens and NetFlix databases. As previously stated, we have used $K = 500$ and the correlation as the similarity measure. As shown, the higher values appear more often in MovieLens than in NetFlix because the sparsity level within NetFlix is greater by far.

Once we have clarified the meaning of $|K_{u,i}|$, we will study the possibility of taking it into account as a factor for measuring the reliability of a prediction $p_{u,i}$. Imagine two predictions, p_{u,i_1}, p_{u,i_2}, such that the following is true:

- About prediction p_{u,i_1}: There are 100 neighbors of u who have rated the item i_1 ($|K_{u,i_1}| = 100$).
- About prediction p_{u,i_2}: Only one neighbor of u has rated the item i_2 ($|K_{u,i_2}| = 1$).

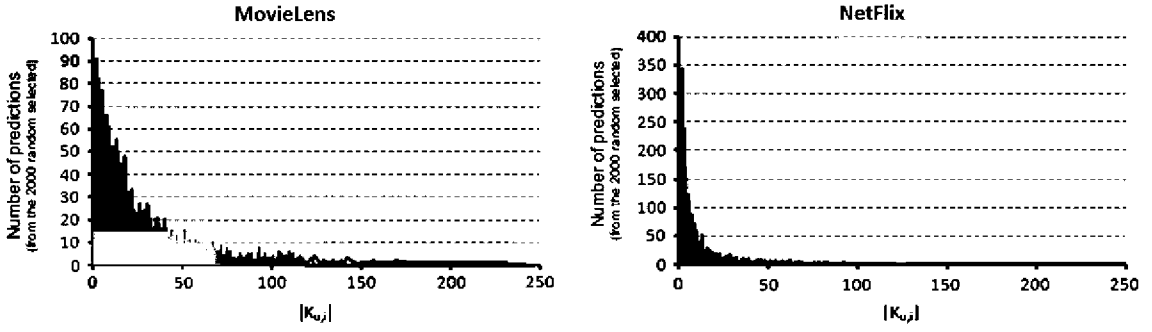

Fig. 2. The distribution of $|K_{u,i}|$ for the MovieLens and NetFlix databases. In these plots, we study the percentage of predictions in the random sample (axis y) for each value in the factor $|K_{u,i}|$ (axis x).

It seems reasonable that the prediction p_{u,i_1} is more reliable than p_{u,i_2} because there are more neighbors of u who have rated the item i_1 than the item i_2. Indeed, while p_{u,i_1} is calculated by taking into account 100 users with similar tastes to u ($|K_{u,i_1}| = 100$), p_{u,i_2} is calculated by taking into account only one user with similar tastes to u ($|K_{u,i_2}| = 1$).

Consequently, we could intuitively state that the greater the value of $|K_{u,i}|$, the more reliable the prediction $p_{u,i}$ (that is to say, $|K_{u,i}|$ may be regarded as a positive factor). However it is only the first step to a more sophisticated factor because it does not take into account how similar the neighbors $K_{u,i}$ are to u. Consider this theoretical case. Two predictions p_{u,i_1} and p_{u,i_2} are such that $|K_{u,i_1}| = |K_{u,i_2}|$, but the similarity between u and every user of K_{u,i_1} is 0.9 and the similarity between u and every user of K_{u,i_2} is 0.01. It seems reasonable to think, that although $|K_{u,i_1}| = |K_{u,i_2}|$, the prediction p_{u,i_1} is much more reliable than p_{u,i_2} because the calculation of p_{u,i_1} involves users who are more reliable than the calculation of p_{u,i_2}.

To include the information related to the similarity of the neighbors $K_{u,i}$, we propose to define the following positive factor:

$$S_{u,i} = \sum_{v \in K_{u,i}} sim(u, v)$$

As shown, the factor $S_{u,i}$ takes into account the number of neighbors who have rated the item i (that is to say $|K_{u,i}|$) as well as the similarity between these users and u.

Once the consideration of the factor $S_{u,i}$ (to measure the reliability of predictions) is justified psychologically, we will try to justify it mathematically as described in Section 4:

In Fig. 3, we show a random sample of 2000 predictions in the *MovieLens* and *NetFlix* databases in which each point placed on the graph $(S_{u,i}, e_{u,i})$ informs the factor $S_{u,i}$ and the real error, $e_{u,i} = r_{u,i} - p_{u,i}$ made in a prediction.

As shown in both plots, the greatest prediction errors are made when the value $S_{u,i}$ has low values, which is exactly what we expected when dealing with a positive factor (see Section 4).

As observed in Section 4, the study of $\text{MAE}_{S_{u,i} \geqslant v}$ is another way to mathematically justify the inclusion of factor $S_{u,i}$ in the reliability measure. Fig. 4 shows the trend of $\text{MAE}_{S_{u,i} \geqslant v}$ in relation to the value v for the *MovieLens* and *NetFlix* recommender systems. As shown, there is a trend in which the higher the value of v, the lower the value of $\text{MAE}_{S_{u,i} \geqslant v}$. Consequently, the value $S_{u,i}$ can be considered as a positive factor in prediction reliability.

Once the use of the factor $S_{u,i}$ is justified, we define the partial reliability of this factor as we have described in Section 4. We have proposed the following function because it fulfills suitable properties (see Proposition 4) for a positive factor:

$$f_S(S_{u,i}) = 1 - \frac{\bar{s}}{\bar{s} + S_{u,i}}$$

where $\bar{s}$ is the median of the values of $S_{u,i}$ in the specific recommender system. In this way, we have defined that $\bar{s} = 26$ for the MovieLens database and $\bar{s} = 5$ for the NetFlix database.

It may be easily proven that the function f_S fulfills Proposition 4, which is related to the definition of a partial reliability of a positive factor (see Definition 2).

Proposition 4. *The following statements hold true:*

(i) If $S_{u,i} = 0$, *then* $f_S(S_{u,i}) = f_S(0) = 0$. *The partial reliability is 0 in the case where there were no neighbors of u who have rated the item i.*

(ii) $\lim_{S_{u,i} \to \infty} f_S(S_{u,i}) = 1$. *The partial reliability is near 1 in the case where there were similar neighbors of u who have rated the item i.*

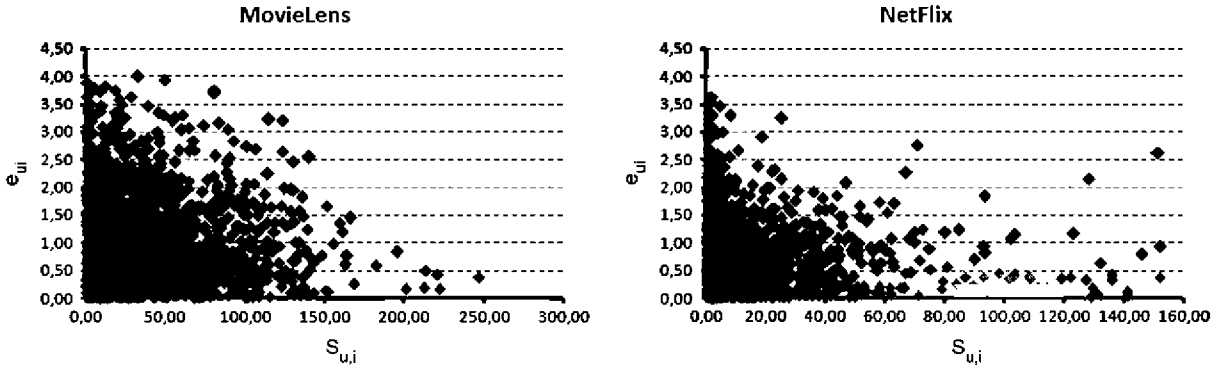

Fig. 3. A scatter plot of the factor $S_{u,i}$ for the MovieLens and NetFlix databases. In these plots, each prediction $p_{u,i}$ is represented by a point where the x-axis stands for the factor $S_{u,i}$ and the y-axis stands for the error made in the prediction $e_{u,i}$.

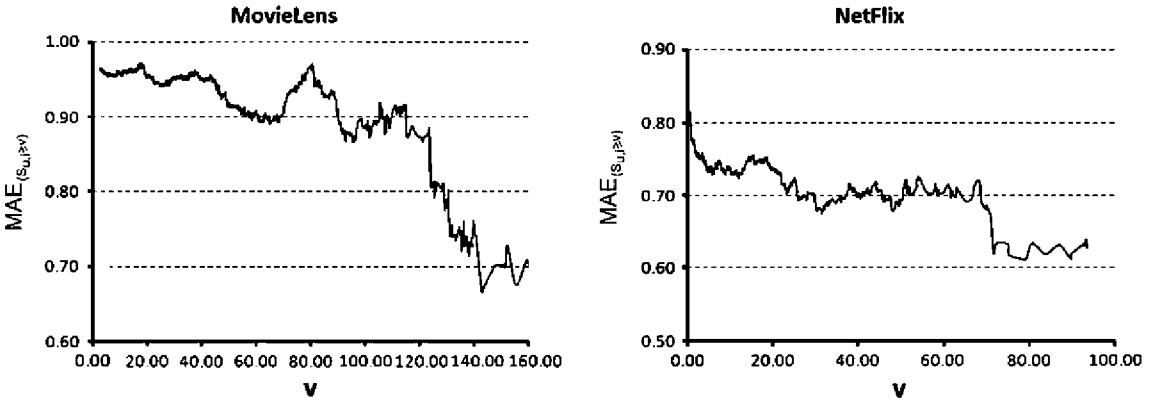

Fig. 4. The $MAE_{S_{u,i}>v}$ for the MovieLens and NetFlix recommender systems. In these plots, we study the mean average error (on *y*-axis) for the predictions in the random sample with a factor $S_{u,i}$ greater than a given value (*x*-axis).

(iii) *If* $S_{u,i} < S_{u',i'}$ *then* $f_S(S_{u,i}) < f_S(S_{u',i'})$ *As we have demanded in a previous section, the partial reliability is an increasing function because* $S_{u,i}$ *is a positive factor of the reliability.*

(iv) *If* $S_{u,i} = \bar{s}$, *then* $f_S(\bar{s}) = 0.5$. *As we have demanded in previous section, the partial reliability is 0.5 when* $S_{u,i}$ *is near to the median of the values of* $S_{u,i}$.

Regarding the importance of the factor $S_{u,i}$, α_S, we believe it should be a constant value for any prediction because it does not depend on any other factor involved. In particular, we propose to consider importance of the factor as:

$$\alpha_S = 1$$

5.2. The factor $V_{u,i}$

In this section, we will study another factor (a negative factor) for measuring the reliability of the prediction of a rating.

Given a prediction $p_{u,i}$, we will consider the variance $V_{u,i}$, of the ratings that the neighbors of the user u have made for the item i, weighted by their similarity. We will consider the following negative factor[6]:

$$V_{u,i} = \frac{\sum_{v \in K_{u,i}} sim(u,v) \cdot (r_{v,i} - \bar{r}_v - p_{u,i} + \bar{r}_u)^2}{\sum_{v \in K_{u,i}} sim(u,v)}$$

As in the previous factor, we will use an example to justify the consideration of this factor for measuring the reliability of the predictions. Consider two predictions, p_{u,i_1} and p_{u,i_2}, such that the following statements are true[7]:

- Prediction p_{u,i_1}. There are 10 neighbors of u who have rated the item i_1, and all of them have rated i_1 with a value of 5. In this prediction, $p_{u,i} = 5$ and the factor $V_{u,i} = 0$.
- Prediction p_{u,i_2}. There are also 10 neighbors of u who have rated the item i_2. All of them have the same similarity as u but five of these neighbors have rated i_2 with a value of 5, and the other five have rated it with a value of 1. In this prediction, $p_{u,i_2} = 3$ and the factor $V_{u,i} = 4$.

It seems reasonable to think that the prediction p_{u,i_1} is more reliable than p_{u,i_2} because there is more agreement between the neighbors of u when rating the item i_1 than when rating the item i_2. Consequently, it seems reasonable to state that the lesser the value of $V_{u,i}$, the more reliable the prediction $p_{u,i}$ (meaning that $V_{u,i}$ is a negative factor according to Section 4).

However, the factor $V_{u,i}$ must be carefully considered. Note that when $|K_{u,i}| = 1$ (consequently, the factor $S_{u,i}$ is likely to also be low), $V_{u,i} = 0$. Indeed, the lower $S_{u,i}$ (the prediction is unreliable as stated in the previous section), the more likely the value $V_{u,i}$ will also be low. Consequently, the factor $V_{u,i}$ must be considered more when measuring the reliability of a prediction in which the value $S_{u,i}$ is high rather than when it is low.

Once the consideration of the factor $V_{u,i}$ is justified psychologically, we will try to justify it mathematically, as described in Section 4:

Of a random sample of 2000 predictions for the *MovieLens* and *NetFlix* databases, Fig. 5 shows only those fulfilling that $f_s(S_{u,i}) \geqslant \alpha$, where α is 0.5, 0.7 and 0.8. Each point placed on the plot on $(V_{u,i}, e_{u,i})$ informs about the factor $V_{u,i}$ and the real

[6] This factor is defined according to the most used aggregation approach described in Section 3.

[7] In this example, we will consider that the mean of the ratings made by the active user, $\bar{r}_u$, is exactly the same as the mean of the ratings made by any of its neighbors, $\bar{r}_v$.

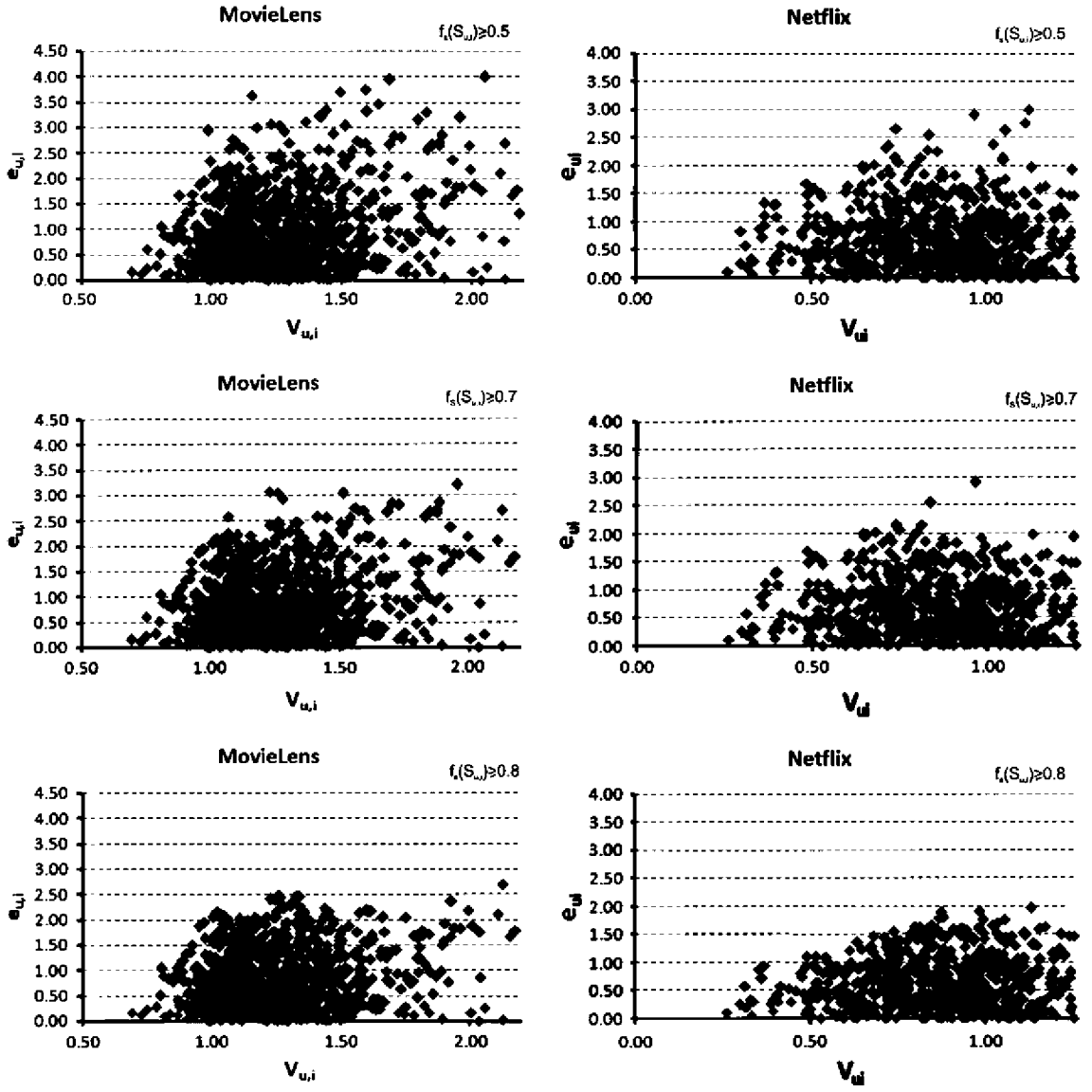

Fig. 5. The scatter plots of the factor $V_{u,i}$ for the MovieLens and NetFlix databases. Each prediction $p_{u,i}$ is represented in these plots by a point where the *x*-axis represents the factor $V_{u,i}$ and the *y*-axis represents the error, $e_{u,i}$, made in the prediction.

error $e_{u,i} = r_{u,i} - p_{u,i}$ made in a prediction. As seen in the plots, the greatest prediction errors are made when the value $V_{u,i}$ takes on high values. That is what we expected when dealing with a negative factor (see Section 4).

In this figure, we can see how the factor $S_{u,i}$ relates to the importance of the factor $V_{u,i}$. As shown, when the value of α is higher (the higher the factor $S_{u,i}$), the error made in these predictions tends to be lower. The higher the factor $S_{u,i}$ value, the more that the factor $V_{u,i}$ must be taken into account.

Now we will study the $\mathrm{MAE}_{V_{u,i} \leq v}$. Fig. 6 shows the value $\mathrm{MAE}_{V_{u,i} \leq v}$ in relation to the value v for the *MovieLens* and *NetFlix* recommender systems. As shown, the higher the value v, the higher the $\mathrm{MAE}_{V_{u,i} \leq v}$. Consequently, the value of $V_{u,i}$ can be considered to be a negative factor as far as prediction reliability is concerned.

Once we have justified the use of the factor $V_{u,i}$, we define the partial reliability of the predictions according to the factor $V_{u,i}$. We have proposed the following function f_V because it fulfills suitable properties (see Section 4) for a negative factor:

$$f_V(V_{u,i}) = \left(\frac{max - min - V_{u,i}}{max - min}\right)^{\gamma}$$

where

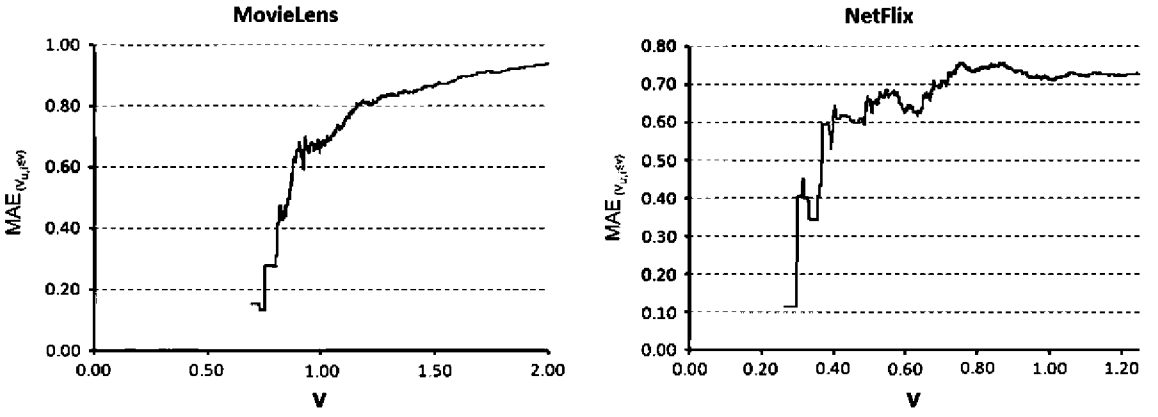

Fig. 6. The $MAE_{V_{u,i}\leq v}$ for the *MovieLens* and *NetFlix* recommender systems. In these plots we study the mean average error (on *y*-axis) for the predictions in the random sample with a factor $V_{u,i}$ that is lesser than a given value (*x*-axis).

$$\gamma = \frac{\ln 0.5}{\ln \frac{max - min - \bar{v}}{max - min}}$$

and $\bar{v}$ is the median of the values of $V_{u,i}$ in the specific recommender system. In this way, we show that $\gamma = 1.98$ for the MovieLens database and $\gamma = 3.55$ for the NetFlix database.

It may be easily proven that, the function f_V fulfills Proposition 5, which is related to the definition of a partial reliability of a negative factor (see Definition 2):

Proposition 5. *The following statement hold true:*

(i) *If* $V_{u,i} = 0$, *then* $f_V(V_{u,i}) = f_V(0) = 1$. *The partial reliability is 1 when there is no disagreement between the neighbors of user u* ($V_{u,i} = 0$).
(ii) *If* $V_{u,i} = max - min$, *then* $f_V(V_{u,i}) = 0$. *The partial reliability is 0 when the maximal disagreement between the neighbors of user u.*
(iii) *If* $V_{u,i} < V_{u',i'}$ *then* $f_V(V_{u,i}) < f_V(V_{u',i'})$. *As we have demanded in previous sections, the partial reliability is a decreasing function because* $V_{u,i}$ *is a negative factor as far as reliability is concerned.*
(iv) *If* $V_{u,i} = \bar{v}$, *then* $f_V(\bar{v}) = 0.5$. *That is to say, such as we have demanded in previous section, the partial reliability is 0.5 when* $V_{u,i}$ *is near to the median of the values of* $V_{u,i}$.

The importance of the factor $V_{u,i}$ is not a constant (unlike the importance of factor $S_{u,i}$). As previously stated, the more important the factor $S_{u,i}$ is for measuring the reliability of predictions, the greater the importance of the factor $V_{u,i}$. Therefore, we propose to calculate α_V (the importance of the factor $V_{u,i}$) as:

$$\alpha_V = f_S(S_{u,i})$$

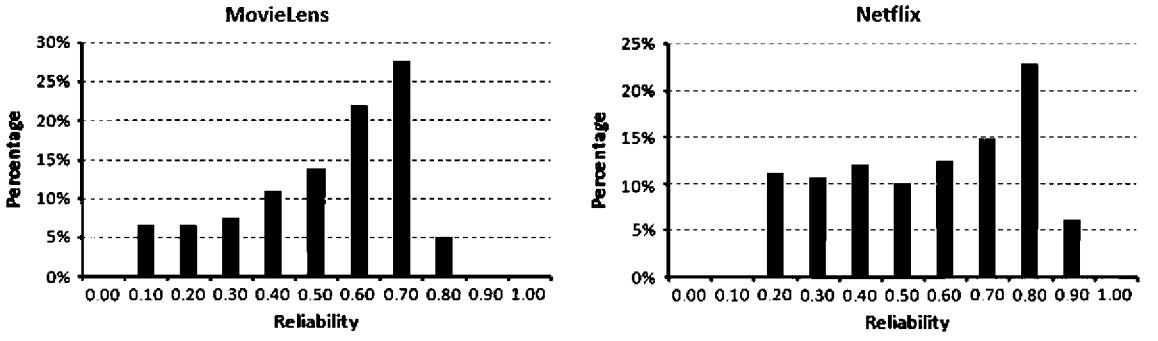

Fig. 7. Distribution of the reliability of predictions for the databases MovieLens and NetFlix. In these plots, we study the percentage of predictions (on *y*-axis) with a given reliability value (*x*-axis).

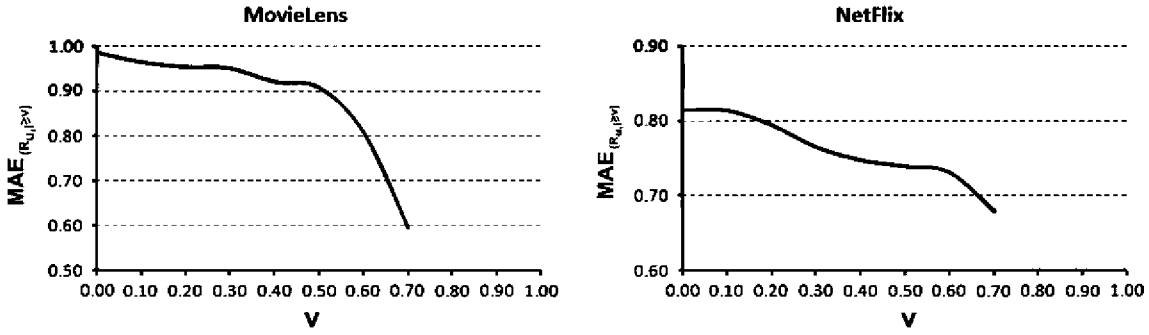

Fig. 8. The $MAE_{R_{u,i}>v}$ for the MovieLens and NetFlix recommender systems. We study the mean average error (on *y*-axis) in these plots for the predictions in the random sample in which the reliability measure is greater than a given value (*x*-axis).

5.3. Calculating the reliability of a prediction

Once the factors $S_{u,i}$ and $V_{u,i}$ are identified, their corresponding partial reliability is obtained, and their relative importance is stated, we can define the reliability of a prediction $p_{u,i}$ by means of Definition 3:

$$R_{u,i} = \left(f_S(S_{u,i}) \cdot f_V(V_{u,i})^{f_S(S_{u,i})}\right)^{\frac{1}{1+f_S(S_{u,i})}}$$

Consider two possible situations related to a prediction $p_{u,i}$, which explains the importance of defining the reliability as a geometric average instead of as an arithmetic average:

(i) There are no neighbors of u who have rated the item i. In this case, $S_{u,i} = 0$; therefore, $f_S(S_{u,i}) = 0$. Regardless of the information provided by other possible factors, it would be reasonable to presume that the prediction is not reliable.
(ii) There are many neighbors of u who have rated the item i, but there is a high disagreement between them when rating the item i. In this case, $V_{u,i}$ is high; therefore, $f_V(V_{u,i}) = 0$. Consequently, we could immediately state that the reliability of the prediction is 0 (despite that the factor $f_s(S_{u,i})$ may be high).

In both cases, the reliability of the prediction $p_{u,i}$ is 0 (as we expected) according to the definition of reliability measure. Note that this would not be possible if the reliability were defined in the form of an arithmetic mean.

6. Results

In this section, we will show how the reliability measure defined in Section 5.3 may be useful for the recommender systems that are based on collaborative filtering. We will study this reliability measure through the *MovieLens* and *NetFlix* recommender systems. As previously stated, we will use the correlation as a similarity measure and a constant $K = 500$.

First, we will show in Fig. 7 the distribution of the reliability measure (defined in Section 5.3) for the *MovieLens* and *NetFlix* recommender systems. As shown, the reliability measure is not concentrated in low values, as we have previously demanded in Section 4. Indeed there are no predictions $p_{u,i}$ with a reliability of 0 ($R_{u,i} = 0$) or a low value, because, as expected, in these databases there are always neighbors who have rated an item ($S_{u,i} > 0$) with a certain agreement between them. In the same manner, there are no predictions, $p_{u,i}$, with a reliability of 1 ($R_{u,i} = 1$) or high value because, as expected, there are no predictions with a high number of neighbors of u who have rated the item i with the same value. Because the reliability measure is merely a subjective measure, we consider it important that the reliability values in a prediction does not fall in the low values. Otherwise, the users would distrust the recommender system.

Now, we will study the relation between the reliability measure and the error in predictions. We will define $MAE_{R_{u,i}>v}$ as the Mean absolute error of those predictions, $p_{u,i}$ which have a reliability measure greater than v (that is to say, $R_{u,i} > v$). In Fig. 8 we depict the relation between $MAE_{R_{u,i}>v}$ and v. As may be observed in this graph (as we expected), the greater v is, the lesser $MAE_{R_{u,i}>v}$ is. That is to say, the reliability measure of a prediction is immediately related to the error as we expected: the more reliable a prediction, the less likely the recommender system will fail in this prediction.

7. Comparing metrics through reliability measure

In the previous section we have studied the reliability of the predictions calculated trough the metric correlation. In this section, we will show how the reliability measure may be used to compare different metrics.

Considering the different metrics, we could study the reliability of the predictions calculated through these metrics, emphasizing the metric that provides the highest prediction reliability.

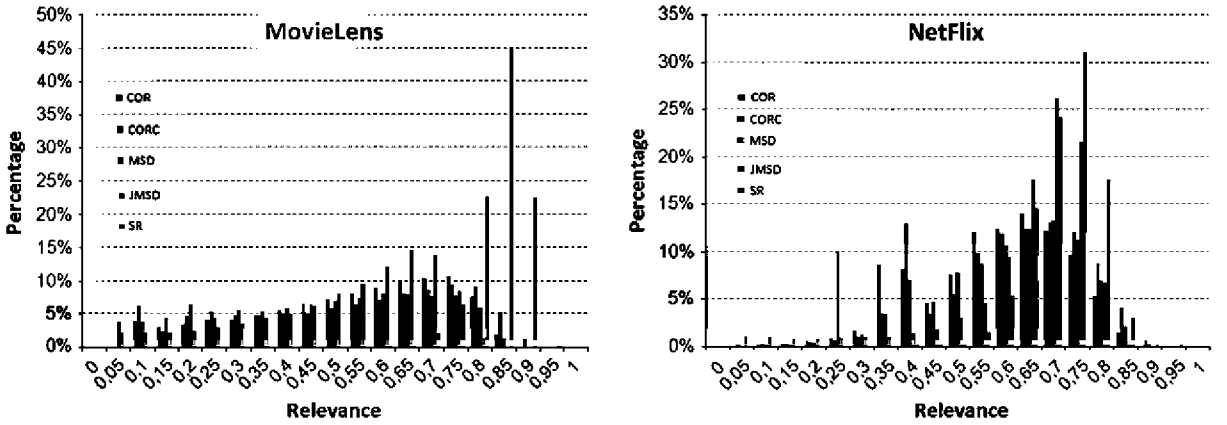

Fig. 9. The reliability of the predictions calculated through different metrics in the MovieLens and NetFlix recommender systems. In these plots we study each metric considered as the percentage (y-axis) of the predictions with a given reliability value (x-axis).

In Fig. 9, we show the distribution of the reliability measure on predictions for the *MovieLens* and *NetFlix* recommender systems using different metrics that are proposed for researchers: Pearson's Correlation (COR); Pearson's Correlation Constrained (CORC); the Mean Squared Difference (MSD); the Jaccard plus MSD (JMSD); and Spearman's Rank-Order Correlation (SR).

As shown, the metric Spearman's Rank-Order Correlation (SR) provides the highest percentage of the most reliable predictions. Most predictions made through this metric are calculated by taking into account a greater number of neighbors (factor $S_{u,i}$) and a wider agreement between these neighbors (factor $V_{u,i}$) than is the case when using any of the other metrics.

8. Conclusions and future work

In this paper we have presented the idea of using a reliability measure associated with the predictions made by a recommender system. In this manner, we will provide a user with a pair of values when recommending an item: a prediction of how much he will like this item; and the reliability measure of this prediction. Using these two values, users could balance between the prediction made by the recommender system and the reliability of this prediction to make their decision.

We have presented a framework for defining reliability measures fitted for a specific recommender system based on collaborative filtering. This framework is based on certain identifying factors that play an important role in the definition of the reliability measure.

We have also defined a reliability measure using two factors that may be suitable for any recommender system based on collaborative filtering. We have tested this reliability measure against real data in two real recommender systems, and we have shown that the more reliable the prediction, the less likely it is that the recommender system fails in this prediction.

As we have stated, this reliability measure may be adapted to a specific recommender system by including new factors (fitting this recommender system). Consequently, future studies could examine (i) new general recommender system factors, particularly the specific factors fitting some recommender systems in particular and (ii) the impact of introducing the pair (prediction, reliability) into the recommendations made to recommender system users, to actually measure the evolution of the user's trust in a recommender system. Another possible opportunity for future work involves defining a new reliability measure by adding specific user-related information, such as demographic, trust and distrust relations and content-based information.

References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (6) (2005) 734–749.

[2] S. Alonso, E. Herrera-Viedma, F. Chiclana, F. Herrera, A web based consensus support system for group decision making problems and incomplete preferences, Information Sciences 180 (23) (2010) 4477–4495.

[3] D. Anand, K.K. Bharadwaj, Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities, Expert Systems with Applications 38 (5) (2011) 5101–5109.