When Users with Preferences Different from Others Get Inaccurate Recommendations

Benjamin Gras(✉), Armelle Brun, and Anne Boyer

LORIA - Université de Lorraine, Vandoeuvre-lès-nancy 54506, France
{benjamin.gras,armelle.brun,anne.boyer}@loria.fr
http://www.loria.fr

Abstract. The social approach in recommender systems relies on the hypothesis that preferences are coherent between users. To recommend a user u some resources, this approach exploits the preferences of other users who have preferences similar to those of u. Although this approach has shown to produce on average high quality recommendations, which makes it the most commonly used approach, some users are not satisfied: they get low quality recommendations. Being able to anticipate if a recommender will provide a given user with inaccurate recommendations, would be a major advantage. Nevertheless, little attention has been paid in the literature to studying this particular point. In this work, we assume that some of the users who are not satisfied do not respect the assumption made by the social approach of recommendation: their preferences are not coherent with those of others; we consider they have atypical preferences. We propose measures to identify these users, upstream of the recommendation process. These measures only exploit the users profile. The experiments conducted on a state of the art corpus and three social recommendation techniques show that the proposed measures allow to identify reliably a subset of users with atypical preferences, who will actually get inaccurate recommendations with a social approach. One of these measures is the most accurate, whatever is the recommendation technique.

Keywords: Atypical preferences · Atypical users · Recommender systems · Collaborative filtering · Accuracy of recommendations

1 Introduction

The continuous increase of the amount of data available on the Internet makes the task of accessing targeted information more and more complex for the users. This is the reason why many services now offer to assist their users during their search, by selecting for them the most relevant information or data. Several types of such services are proposed, among which recommender systems (RSs) [1]. Through a recommendation process, a RS aims to guide a user, called the *active user*: the user the system aims to provide with recommendations, towards resources relevant for him/her. A resource can be a book, a movie, a

V. Monfort et al. (Eds.): WEBIST 2015, LNBIP 246, pp. 191–210, 2016.
DOI: 10.1007/978-3-319-30996-5_10

web page, etc. To make such a recommendation possible, the system uses the knowledge it has collected about this active user.

RSs have been studied for more than twenty years [1]. The two most common approaches are content-based filtering [2] and collaborative filtering (CF) [3,4]. Content-based filtering exploits the content of the resources (as well as indexes, keywords, title, type of the resource, etc.) to select those that match the active user's preferences. Conversely, CF (also referred to as social filtering) does not require the exploitation of the content of the resources. It relies on the assumption that users' preferences are consistent among users, which allows to infer the active user's preferences from those of other users. In both approaches, users' preferences are generally represented by ratings on the resources. As CF is the most popular approach, it will be the focus of this work.

Providing users with high quality recommendations is of the highest importance. Indeed, in the context of e-commerce it increases customer retention, in e-learning it improves learners' learning process, in digital libraries it allows users to save time, etc. The quality of the recommendations provided by CF is now considered as acceptable on average [5]. However, some users do not receive accurate recommendations, which results in serious consequences: unsatisfied users, customer attrition, failure among learners, time wasted, etc.

If we are unable to provide each user with accurate recommendations, we are convinced it is essential that a given recommender can anticipate, upstream of the recommandation process, the users it will provide with inaccurate recommendations. Once these users are identified, the system can decide to not provide them with recommendations at all, or decide to use another approach specifically dedicated to these users. The literature has emphasized that one reason why some users are not satisfied is the small number of preferences the system collected about them. This problem is referred to as the cold-start problem [6]. However, some users with a significant number of preferences still get inaccurate recommendations. This can also be explained by the quality of the preferences collected about these users [7] or by the inconsistency of the users when expressing their preferences [8]. Recent works have noticed that some specific users tend to rate resources differently than other users [9,24]. We will refer their preferences to as atypical preferences. Remind that collaborative filtering assumes that preferences (ratings) are consistent between users. As these users do not match this requirement (their preferences are not consistent with those of others), this may explain why some of them get inaccurate recommendations.

The work conducted in this paper is in line with these latter works. We aim at identifying reliably users with atypical preferences (ratings) and who will receive inaccurate recommendations. From now on, we will refer these users to as atypical users. Their identification will be performed prior to any recommendation computation. To reach this goal, we introduce several measures that reflect the atypicity of preferences of a user.

Section 2 presents a short overview of recommender systems and the way atypical users are identified and managed in social recommendation. Section 3 introduces the three measures we propose to identify atypical users. Then, in

Sect. 4 the experiments we conducted to evaluate those measures are presented. Finally, we conclude and discuss our work in the last section.

2 Related Works

2.1 Social Recommender Systems

To provide a user, referred to as the *active user*, with some personalized recommendations, the social recommendation, also denoted by collaborative filtering (CF) [3,4], relies on the knowledge of other users preferences (generally some ratings) on resources. When the ratings are not available, preferences can be inferred from the traces of activity left by the users [10].

There are two main techniques in social recommendation: the memory-based technique and the model-based technique [11]. The memory-based technique (also referred to as instance-based learning) exploits directly users' preferences, without pre-processing. The most commonly used technique, the K Nearest Neighbors (KNN) user-based paradigm [3], exploits neighbor users of the active user. First, it computes the similarities of preferences between the active user and each other user. There are many ways to compute the similarities, the most popular is the Pearson correlation coefficient presented in Eq. (1). Second, it identifies the k nearest neighbors of u who have rated r(those with the highest similarity value). Last, it computes an estimation of the active user's rating using the ratings of his/her K nearest neighbors, using the weighted mean average (see Eq. (2)).

$$Pearson(u,v) = \frac{\sum_{r \in R_{uv}} (n_{u,r} - \overline{n}_u)(n_{v,r} - \overline{n}_v)}{\sqrt{\sum_{r \in R_{uv}} (n_{u,r} - \overline{n}_u)^2} \sqrt{\sum_{r \in R_{uv}} (n_{v,r} - \overline{n}_v)^2}} \tag{1}$$

where $n_{u,r}$ is the rating of the user u on the resource r, R_{uv} is the set of co-rated resources by users u and v and $\overline{n}_u$ is the average rating of u.

$$n^*_{u,r} = \overline{n}_u + \frac{\sum_{v \in V_{u,r}} (n_{v,r} - \overline{n}_v) * sim(u,v)}{\sum_{v \in V_{u,r}} |sim(u,v)|} \tag{2}$$

where $n^*_{u,r}$ is the estimated rating of user u on resource r, $V_{u,r}$ represents the k nearest neighbors of u, who rated the resource r and *sim(u,v)* is the similarity calculated between u an his/her neighbor v. The similarity can be instantiated by the Pearson correlation coefficient (see Eq. (1)). r Another well-known memory-based paradigm is the item-based paradigm which computes the similarities between items (resources) to deduce preferences of users. This paradigm relies on the hypothesis that if a user like a resource r, he/she will like the most similar resources to r. Once more the Pearson correlation coefficient (presented in Eq. (1)) is the most popular similarity measure used in the item-based paradigm.

The memory-based technique is simple to implement, provides high quality recommendations and takes into account each new preference dynamically in the recommendation process. However, it does not scale, due to the computation cost of the high number of similarities.

The model-based technique learns, as its name suggests, a model that describes the data (preferences). This model is used to estimate unknown preferences, so to provide the active user with recommendations. This approach does not suffer so much from the scalability problem. However, it does not easily allow dynamic changes in the model, especially if it has to be updated each time a new preference is provided by a user.

The model-based matrix factorization technique [12] is now the most commonly used technique, due to the quality of recommendations it provides. The matrix of users' preferences is factorized into two sub-matrices, one representing users, the other representing the resources, both in a common sub-space where dimensions correspond to latent features. Then, to compute a estimation of the rating of the user u on the resource r, the recommender multiply the vector of latent features associated to u by the vector of latent features associated to r. There are several matrix factorization techniques, including the singular value decomposition (SVD) [13] and alternating least squares (ALS) [14].

One limit, common to all CF techniques (whether memory-based or model-based) is the cold-start problem [6], which is related to the lack of data on new resources or new users.

2.2 Identifying Atypical Users in Recommender Systems

In the literature, several terms are used to make reference to atypical users. They are deviant users [9], abnormal users [15], grey sheeps [16], etc. T Most of the techniques used to perform their identification are issued from data analysis. The abnormality measure [9,15] is the most commonly used one. It has actually several names such as abnormality or deviance. Those names reflect the tendency of a user to rate differently from others. This measure exploits the difference between the ratings assigned by a user on some resources and the average rating on these resources. It is defined by Eq. (3).

$$Abnormality(u) = \frac{\sum_{r \in R_u} |n_{u,r} - \overline{n_r}|}{||R_u||} \tag{3}$$

where $n_{u,r}$ represents the rating that user u assigned to resource r, $\overline{n_r}$ is the average rating of r among all users, R_u is the set of resources rated by u and $||R_u||$ is their number. The higher a user rates resources differently than the average user, the higher his/her abnormality value. Users with a high abnormality value are considered as atypical users. The main advantage of this measure is its low complexity. However, although it is the reference measure in the literature to identify users with atypical preferences, from our point of view it suffers from several limitations. First, the resources about which users' preferences are not unanimous (the ratings between users is very different) will unfairly increase the abnormality of the users who rate these resources. Second, this measure does not take into account the individual behavior of each user. For example, a user more strict than the average user may be labeled as abnormal, while he/she has similar preferences to others; he/she only differs in his/her way of rating resources. This measure will thus probably identify some users as atypical, whereas they will get accurate recommendations.

Some studies identify atypical users with the aim to explain the fluctuations of performance of RS [15,17–19]. To reach this goal, they study users' characteristics: number of ratings, number of neighbors, etc. For example, a link between the small number of ratings of a user and a high recommendation error may be identified (cold-start problem). In [15], the authors form clusters of users, based on their preferences and aim at interpreting the resulting clusters. Among the set of clusters, a cluster made up of atypical users is identified: users with a high recommendation error (RMSE) as well as a high abnormality (Eq. (3)) as well. However, we are convinced that in the general case, clustering fails to build a cluster of users with atypical preferences and who will get inaccurate recommendations. Indeed, an atypical user, in the sense of the social recommendation, has preferences that are not close to those of other users. Thus, if a user belongs to a cluster, it means that his/her preferences are similar to those of users in the same cluster. So, he/she is not an atypical user. The work presented in [16] also relies on clustering of users, and is in line with our conviction: it proposes to consider users who are far from the center of their cluster as atypical users.

[17] defines a clarity indicator, that represents how much a user is non-ambiguous in his/her ratings. This indicator is based on the entropy measure: a user is considered as ambiguous (small value of clarity) if his/her ratings are not stable across resources. Authors show that there is a link between the ambiguity of the ratings of a user and the quality of recommendations he/she gets. Users with a small clarity value are considered as noise and are discarded from the system; they do not receive any recommendations. We believe that this approach quickly appears constrained. Indeed, various ratings (preferences) of a user can be explained by several factors such as the evolution of his/her preferences through time, his/her varying preferences across domains, etc. Therefore, a social approach may anyway provide this user with high quality recommendations. Notice that, at the opposite of previous approaches, the clarity indicator does not reflect the coherence of a user's preferences with respect to other users, it reflects the coherence he/she has with him/herself. It can thus be exploited in an approach other than the social one. Clarity can also be linked the magic barrier concept [20] and to recent works about user inconsistency and natural variability [21], which aim at estimating an upper bound on the rating prediction accuracy.

The impact of users identified as atypical on the overall quality of recommendations has been studied. The comparison of the results presented is difficult as atypical users are not selected on the basis of the same criteria. However, they do all conclude that removing atypical users in the learning phase of the recommender improves the overall quality of the recommendations.

Notice that the identification of atypical users may be associated with the identification of outliers or anomalies. According to [22], an outlier is "an observation that deviates so much from other observations as to arouse suspicion that is was generated by a different mechanism". In the context of recommender systems, an outlier is a user whose preferences appear to have been generated by a different preference expression mechanism. Criterion based, statistical approaches and clustering are also widely used in the field of outliers detection [23].

2.3 Managing Atypical Users in Recommender Systems

Once atypical users have been identified, one question that can be addressed is related to their management. In the context of recommender systems, new recommendation approaches dedicated to these specific profiles have been proposed, with the aim to provide them with better recommendations.

In [9], which refers atypical users to as deviant users, the authors divide the set of users into two subsets: deviant and non-deviant users, using the abnormality measure (Eq. (3)). These two subsets are considered independently when training recommendation models (two models are formed), as well as during the recommendation process. Only deviant users are taken into account when the active user is identified as deviant. Conversely, only non-deviant users are considered when the active user is non-deviant. This approach has shown to improve the quality of recommendations provided to non-deviant users. However, it has no impact on the quality of the recommendations provided to deviant users. This confirms our intuition that atypical users do not share preferences with any other user. In addition, we find this result not surprising as the recommendation approach has not been adapted to these specific users.

We previously reported how [16] identify atypical users through clustering. To address these atypical users, they use a specific cluster-based CF algorithm (model-based approach) to better reflect the preferences of these users and to offer them more accurate recommendations. Authors assume that these users only have partial agreement with the rest of the community (i.e. CF will fail on these users) and propose to rely on the content of resources to generate recommendations.

Finally, J. Bobadilla [24] has proposed a more general solution to take into account the specificities of atypical users, through a new similarity measure. This new measure is based on the singularity of ratings. A rating on a resource is considered as singular if it does not correspond to the majority rating on this resource. Authors assume that atypical users tend to assign singular ratings to resources. The singularity is used when computing the similarity between users: the more a rating is singular, the greater is its importance. The similarity between users is then used as in a classical KNN user-based recommendation approach. It has shown to provide high quality recommendations to users with specific preferences.

3 New Atypical Users Identification Measures

In this section, we introduce new measures for identifying atypical users, *i.e.* users with preferences that differ from those of the population of users.We consider that an atypical user receives inaccurate recommendations. These identification measures are designed to be used prior to the recommendation process, so they only rely on the users' profiles (preferences on resources). We want to propose measures that wont select any user receiving accurate recommendations, to not have a negative effect on him/her.

3.1 CorrKMax

The first measure we propose is dedicated to the user-based *KNN* technique. We are convinced that the user-based approach, which exploits the K most similar users to the active user, fails in the case the active user does not have enough highly similar users. We thus define $CorrKMax$ to highlight the link between the similarity of the most similar users of a user u and the quality of the recommendations he/she gets. $CorrKMax(u)$ (Eq. (4)) represents the average similarity between the active user u and his/her K most similar users.

$$CorrKMax(u) = \frac{\sum_{v \in Neigh(u)} Pearson(u,v)}{||Neigh(u)||} \tag{4}$$

where $Pearson(u,v)$ is the Pearson correlation between the preferences of users u and v (see Eq. (1)). $Neigh(u)$ represents the k most similar users to u, in the limit their correlation with u is positive. We believe that the users associated with a low value of $CorrKMax(u)$ receive inaccurate recommendations.

The two following measures are an extension of the *Abnormality* measure from the state of the art, which has shown good atypical users identification capabilities (see Sect. 2.2). To overcome the limitations that we have mentioned and presented in the previous section, we propose a first improvement.

3.2 AbnormalityCR

The *AbnormalityCR* (Abnormality with Controversy on Resources) measure assumes that the meaning of the discrepancy between a rating on a resource and the average rating on this resource differs according to the resource. Indeed, a large discrepancy on a controversial resource has not the same meaning as a large discrepancy on a consensual resource. The abnormality measure of the state of the art considers these differences as equal, which has the effect of increasing the abnormality of users who express their preferences on controversial resources. We therefore propose to reduce the impact of the ratings on controversial resources, by weighting them with the degree of controversy of the resources they refer to.

This degree of controversy of a resource is based on the standard deviation of the ratings on this resource. The *AbnormalityCR* of a user u is computed as shown in Eq. (5).

$$Abnormality_{CR}(u) = \frac{\sum_{r \in R_u} ((n_{u,r} - \overline{n_r}) * contr(r))^2}{\|R_u\|} \tag{5}$$

where *contr*(r) represents the controversy associated with resource r. It is based on the normalized standard deviation of ratings on r and is computed according to Eq. (6).

$$contr(r) = 1 - \frac{\sigma_r - \sigma_{min}}{\sigma_{max} - \sigma_{min}} \tag{6}$$

where σ_r is the standard deviation of the ratings associated with the resource r. σ_{min} and σ_{max} are respectively the smallest and the largest possible standard deviation values, among resources. The computation complexity of *AbnormalityCR* is comparable to that of the abnormality of the state of the art. It can therefore be computed frequently and thus take into account new preferences.

3.3 AbnormalityCRU

The *AbnormalityCRU* (Abnormality with Controversy on Resources and Users) measure is a second improvement of the *Abnormality* measure. It starts from the observation that neither *Abnormality*(u) nor *AbnormalityCR*(u) reflect the general behavior of the user u. Thus, with these measures, a user who is strict in his/her way to rate resources may be considered as atypical, even if his/her preferences are actually not. In addition, this user will probably receive high quality recommendations. To avoid this bias, we propose to center the ratings of each user around his/her average rating. This way to reflect the user's behavior is also the one used in the Pearson correlation coefficient. Furthermore, the average rating on a resource is computed on the centered ratings, as well as the controversy. The abnormality of a user u, denoted by *AbnormalityCRU*(u), is computed using Eq. (7).

$$Abnormality_{CRU}(u) = \frac{\sum_{r \in R_u} [(|n_{u,r} - \overline{n_u} - \overline{n_{C_r}}|) * contr_C(r)]^2}{\|R_u\|} \tag{7}$$

where $\overline{n_{C_r}}$ represents the average centered rating on the resource r, $contr_C(r)$ represents the controversy associated with resource r, computed from the standard deviation of the ratings on u, centered with respect to users. The computation of *AbnormalityCRU*(u) is more complex than *AbnormalityCR*(u), but should allow a more accurate identification of atypical users.

Note that these last two measures are independent of the recommendation technique used, whether it is KNN or matrix factorization, contrary to the $CorrKMax$ measure, dedicated to the user-based KNN technique.

4 Experiments

The experiments we conduct in this section are intended to assess the quality of the atypical users identification measures we propose ($CorrKMax$, *AbnormalityCR* and *AbnormalityCRU*) in comparison with the measure from the state of the art (*Abnormality*). The assessment is based on the quality of the recommendations, more precisely the errors, provided to users identified as atypical.

4.1 Errors Measures

The quality of recommendations is evaluated through two standard measures: the Root Mean Square Error (RMSE) and the Precision. The former exploits the discrepancy between the rating provided by a user on a resource and the rating estimated by the recommender, the latter corresponds to the proportion of accurate predictions. The lower the RMSE, the higher the accuracy of recommendations provided to users. On the contrary, the higher the Precision, the higher the accuracy of the recommender system. In this work, we will specifically exploit per-user RMSE ($RMSE(u)$) and Precision ($Precision(u)$), computed respectively by Eqs. (8) and (9).

$$RMSE(u) = \sqrt{\frac{\sum_{r \in R_u} (n_{u,r} - n^*_{u,r})^2}{||R_u||}} \tag{8}$$

$$Precision(u) = \frac{||(|n_{u,r} - n^*_{u,r}| < 0.5)||}{||R_u||} \tag{9}$$

where $n^*_{u,r}$ is the estimated rating of user u on resource r.

4.2 Dataset and System Settings

Experiments are conducted on the MovieLens100K[1] dataset from the state of the art. MovieLens100K is made up of 100,000 ratings from 943 users on 1,682 movies (resources). The ratings range from 1 to 5, on integer values. We divide the dataset into two sub-sets made up of 80 % (for learning) and 20 % (for test) of the ratings of each user.

As presented in the beginning of this paper, our goal is to identify the users who will be provided with inaccurate recommendations, due to their atypical preferences. The literature emphasizes that users about who the system has collected few preferences get inaccurate recommendations (cold-start problem). To not bias our evaluation, we decide to discard these users from the dataset. We consider that a user who has less than 20 ratings in the training set is associated to cold-start [25]. The set of users is then reduced to 821 users (97 k ratings).

To compute the per-user errors, we implement three different commonly used CF techniques: a user-based technique, an item-based technique and a matrix factorization technique (see Sect. 2.1). Evaluating the atypical users identification measures on various techniques will allow us to determine which measure fits which technique or if these measures are generic: they are accurate whatever is the technique used. We set up the mostly used settings in the state of the art for each of these three techniques.

The user-based technique defines the similarity of two users as the Pearson correlation coefficient (see Eq. (1)) between their two rating vectors. The rating

[1] http://grouplens.org/datasets/movielens/.

estimation for a user is based on the ratings of his K nearest neighbors, using a weighted average of their ratings (see Eq. (2)). We fix K = 20 for this dataset.

The item-based technique defines the similarity between two items as the Pearson correlation coefficient (see Eq. (1)) between their rating vectors. The rating estimation for a user u on an resource r is based on the most similar items to r rated by u, we used a weighted average of the ratings. Such as in the user-based recommender, we fix the number of most similar items to K = 20.

We use the ALS factorization technique to compute the matrix factorization with 5 latent features. The ALS factorization is the most accurate technique to manage sparse matrices.

In order to give us a first overview of the link between those two elements, in the following section, we focus on the correlation between errors calculated with those techniques and our identification measures.

4.3 Correlations Between Identification Measures and Recommendation Error

Four measures are studied in this section: *Abnormality* from the state of the art and the three measures we propose: *AbnormalityCR*, *AbnormalityCRU* and *CorrKMax* (with $K = 20$). Based on these correlations, we can determine which measures are good indicators of the quality of recommendations that will be proposed to users.

The correlations on the user-based recommender are presented in Table 1.

Table 1. Correlations between identification measures and RMSE/Precision of a user-based technique.

	RMSE	Precision
Abnormality	0.453	−0.274
AbnormalityCR	0.504	−0.305
AbnormalityCRU	0.546	−0.364
CorrKMax	−0.22	0.07

Let us first focus on the *Abnormality* measure from the state of the art. Its correlation with RMSE is 0.453. This correlation is significant and confirms the existence of a link between the *Abnormality* of a user and the accuracy of the recommendation he/she gets: the higher the *Abnormality* of a user, the higher the error made on the rating estimation, so the lower the accuracy of the recommendations he/she receives. At the opposite, the lower the *Abnormality*, the higher the accuracy. Recall that a user with a high *Abnormality* value is considered as atypical. The correlation between *Abnormality* and Precision is less significant (−0.274) but does not negate the previous conclusions.

When considering *AbnormalityCR*, the correlation with RMSE reaches 0.504, which corresponds to an improvement of 11 % of the correlation compared to *Abnormality*. In parallel, the correlation of the *AbnormalityCR* with the Precision is also improved by 11 % compared to *Abnormality*. We can deduce that integrating the controversy associated with the resources in the computation of the Abnormality improves the estimation of the accuracy of the recommendations provided to users.

With *AbnormalityCRU*, the correlation with RMSE is equal to 0.546, which corresponds to a further improvement of 8 % (20 % with respect to *Abnormality*) and the correlation with Precision is equal to 0.364, which correspond to a further improvement of 19 % (32 % with respect to *Abnormality*). So, taking into account users' rating peculiarities (users' average rating) further improves the estimation of the accuracy of recommendations.

The correlation between *CorrKMax* and RMSE (−0.22) or Precision (0.07) indicates that, contrary to our intuition, the quality of a user's neighborhood is not correlated with the quality of the recommendations provided to him/her, with a *KNN* recommendation technique. This result is surprising as the *KNN* technique assumes that the more a user is correlated with the active user, the more he/she is reliable, and thus the more important he/she is in the computation of recommendations for this active user.

Table 2 presents the correlations on the item-based recommender.

Table 2. Correlations between identification measures and RMSE/Precision of a item-based technique.

	RMSE	Precision
Abnormality	0.398	−0.225
AbnormalityCR	0.421	−0.252
AbnormalityCRU	0.480	−0.363
CorrKMax	−0.09	0.03

With the item-based technique, the correlations are all weaker than with the user-based technique. Nevertheless, most of those correlations are still significant such as the correlation between *Abnormality* and RMSE with a value of 0.398. The *AbnormalityCR* measure increases this correlation to 0.421 (+6 %) and the *AbnormalityCRU* measure increases it to 0.480 (+20 %). Similar improvements can be measured on the Precision. The correlation between *Abnormality* and Precision is equal to −0.225. The correlation increases by 12 % with *AbnormalityCR* (−0.252) and increases by 61 % with *AbnormalityCRU* (−0.363).

We can then conclude about memory-based approaches that *AbnormalityCR* and *AbnormalityCRU* add some important and useful information to the state of the art *Abnormality*. The *AbnormalityCRU* measure is once more the more correlated with the errors. The *CorrKMax* is absolutely not tied to the errors of the item-based technique, even less than with the user-based technique.

The correlations on the matrix factorization technique are presented in Table 3.

Table 3. Correlations between identification measures and RMSE/Precision of a matrix factorization technique.

	RMSE	Precision
Abnormality	0.432	−0.297
AbnormalityCR	0.409	−0.285
AbnormalityCRU	0.488	−0.398
CorrKMax	−0.20	0.15

The correlation between the RMSE and the *Abnormality* measure is equal to 0.432 and the correlation between RMSE and *AbnormalityCR* is only equal to 0.409. This means that *Abnormality* is more related to the RMSE of a matrix factorization recommender than *AbnormalityCR*. This could indicate that the matrix factorization process can reduce the impact of the controversy of resources on errors. Furthermore, as on the memory-based techniques, the *AbnormalityCR* measure is less correlated with the Precision on matrix factorization errors. The *AbnormalityCRU* measure is once again the more correlated measure with both errors. It improves the correlation between the RMSE and *Abnormality* from 0.432 to 0.488 (+13 %). The *CorrKMax* measure shows its best correlations with the matrix factorization technique. However, the correlations are still not significant enough (−0.20).

In conclusion, with the three techniques, *AbnormalityCRU* is the more related measure to the system errors. At the opposite, the *CorrKMax* is correlated to none of those three techniques errors. Another remark is that the four studied measures are more correlated with the user-based errors.

4.4 Recommendation Error for Atypical Users

The correlations studied in the previous experiments aimed at evaluating the relationship between the abnormality measures and the recommendation errors on the complete set of users. However, there may be a relationship within only a subset of users. In that case, the correlation may not allow to identify this relationship. In particular, in this paper we aim at identifying a link between users identified as atypical and error measures. Therefore, in the following experiments, we will no more focus on the correlation between identification measures and errors measures, but only on the errors observed on users identified as atypical. The users with an extreme value of the identification measure are considered as atypical (the highest ones for the abnormality measures).

To study these errors, we depict them with the minimum, the maximum, the quartiles and the median values, and draw box plots. The four identification measures: *Abnormality*, *AbnormalityCR*, *AbnormalityCRU* and *CorrKMax* are studied.

To evaluate precisely these four measures, we compare their box plots with the one of the complete set of users (denoted by Complete in Figs. 1, 2, 3 and 4).

Recall that, the higher the RMSE, the more accurate the measure and the lower the Precision, the more accurate the measure. As the identification measures do not all have comparable values, we did not use a predefined atypicity threshold value. We chose to consider a predetermined percentage of atypical users, which we fixed experimentally at 6 % of the complete set of users. This corresponds to about 50 users among the 821 users. We compare these measures in the framework of the three recommendation techniques: the user-based technique, the item-based technique and the matrix factorization technique.

Errors Associated with Atypical Users in the User-Based Technique. The distribution of the errors (RMSE and Precision) obtained with the user-based technique, according to the identification measure, are presented in Figs. 1 and 2 respectively.

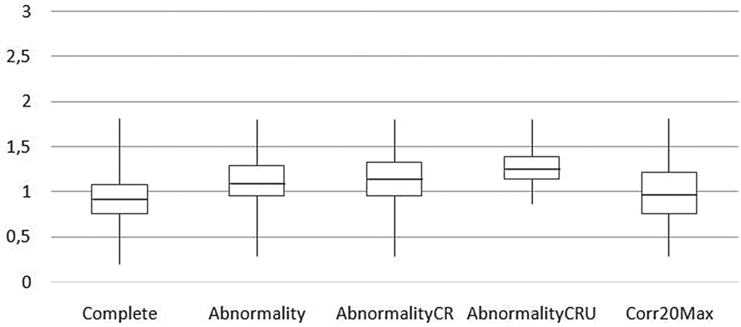

Fig. 1. Distribution of RMSE of atypical users with the user-based technique.

The median RMSE on the complete set of users (Complete) is 0.91. When exploiting the *Abnormality* measure, the median RMSE of the 6 % users with the highest *Abnormality* reaches 1.12. This represents an increase in the RMSE by more than 25 %. Furthermore, we can notice that the median value of *Abnormality* is equal to the third quartile of the Complete set. This mean that 50 % of users identified as atypical users with *Abnormality* are part of the 25 % of users with the highest RMSE in the Complete set: this measure is quite accurate. However, 25 % of the users considered as atypical have a RMSE lower than the median RMSE of the complete set of users. This means that, although *Abnormality* from the state of the art allows to identify users who will receive inaccurate recommendations, it appears to select a significant number of users who will receive accurate recommendations (false detection). *Abnormality* is thus not precise enough. Recall that users identified as atypical may either not receive any recommendations at all, or may get recommendations from another technique, which may be less accurate. It is very important to not identify users as atypical if they will receive high quality recommendations in order to not modify their recommendations. The accuracy of the measure used is thus of the highest importance. The limits of the *Abnormality* measure that we presented in the previous section (see Sect. 2.2) are confirmed: the use of the discrepancy between a rating and the average rating on a resource is not sufficient to reliably predict inaccurate recommendations.

The quality of both *AbnormalityCR* and *AbnormalityCRU* measures is higher than the one of *Abnormality*. *AbnormalityCR* slightly improves the performance of the *Abnormality* measure with a median equal to 1.17 (increase of 4 %). *AbnormalityCRU* appears to be the best one: all the users identified as atypical users have a RMSE higher than the median RMSE of the complete set of users. In addition, over 75 % of these users have a RMSE higher than 1.13, *i.e.* 75 % of the users with the highest *AbnormalityCRU* are among the 25 % of the complete set of users who will receive inaccurate recommendations. The accuracy of the *AbnormalityCRU* measure is thus high.

Once more *CorrKMax* (with $K = 20$) is not accurate, the users identified as atypical tend to receive high quality recommendations (50 % of them). The low similarity of a user's nearest neighbors is thus not a reliable information to predict the low quality of recommendations this user will receive.

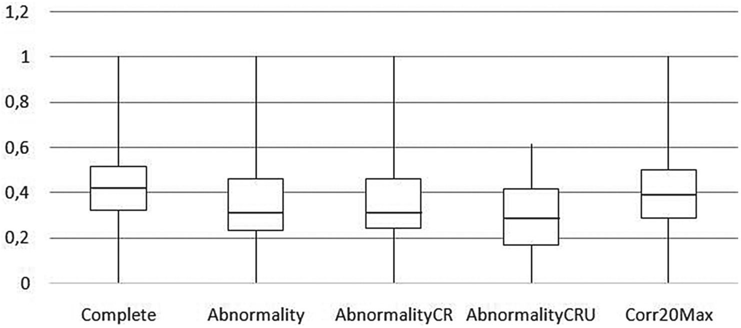

Fig. 2. Distribution of Precision of atypical users with the user-based technique.

The distributions of the Precision, presented in Fig. 2, confirm the results obtained with RMSE. The median Precision obtained on the complete set of users is equal to 0.42. The median Precision of *Abnormality* and *AbnormalityCR* is 0.31, which correspond to an improvement of 35 %. Moreover, 25 % of the complete set of users obtain a Precision lower than 0.32, which mean that 50 % of users selected with the *Abnormality* and *AbnormalityCR* measures belong to the set of 25 % worst Precisions of the system. In contrast to RMSE, we can observe that, with the Precision measure, *AbnormalityCR* does not improve the performance of *Abnormality*. Nevertheless, those measures select also users with accurate recommendations. The median Precision obtained with *AbnormalityCRU* is 0.28, which is not significantly lower than the median Precision of *Abnormality*, but we can see that *AbnormalityCRU* does not select users receiving accurate recommendations. *AbnormalityCRU* is thus the better measure to select users receiving inaccurate recommendations.

The results obtained with $CorrKMax$ (nearly a random selection) are, once more, not conclusive.

The results obtained with Precision are less clear-cut than those obtained with RMSE on those four measures. We can thus deduce that the controversy on resources is more effective at aiming high range deviations between estimations and users ratings than low range deviations. We can conclude that, when the $AbnormalityCRU$ measure identifies a user as an atypical user, he/she will actually receive inaccurate recommendations with the user-based recommendation technique.

Errors Associated with Atypical Users in the Item-Based Technique. The distribution of the errors obtained with the item-based technique are presented in Figs. 3 and 4.

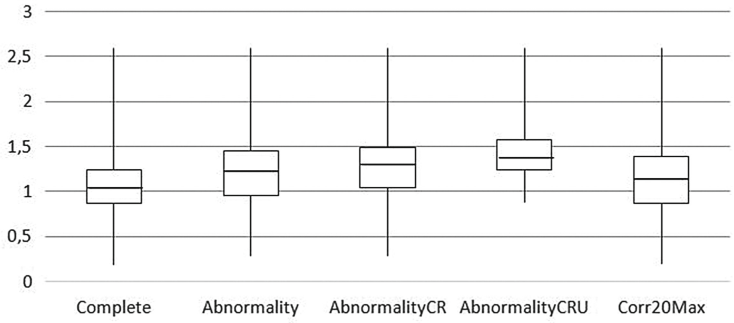

Fig. 3. Distribution of RMSE of atypical users with the item-based KNN technique.

With the item-based technique, the median RMSE of the complete set of users states at 1.04 and 25 % of users have a RMSE higher than 1.23. The median RMSE of the users select with *Abnormality* is equal to 1.27, which corresponds to an increase of the median RMSE by 22 % and means that 50 % of users selected with *Abnormality* belong to the 25 % of users of the Complete set with the worst RMSE. As with the user-based technique, those results are successively increased with $AbnormalityCR$ and $AbnormalityCRU$. The best results are once more obtained with $AbnormalityCRU$: 75 % of users have a RMSE within the 25 % RMSE of the system. The conclusions about this technique are the same than those obtained with the user-based technique.

According to the Precision measure, *AbnormalityCRU* shows, once more, the best results: 75 % of the users selected belong to the set of the 25 % worst RMSE in the complete set of users. In Figs. 3 and 4, the distributions of the errors associated with $CorrKMax$ are once more not conclusive. We can then conclude that in a memory based approach (user-based or item-based), the 20 highest correlations between items or users are not enough to predict the quality of recommendations.

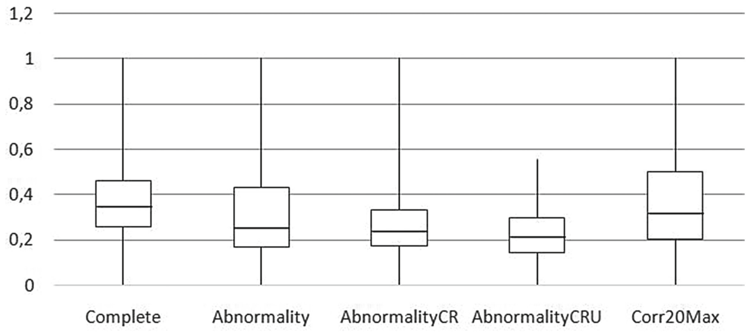

Fig. 4. Distribution of Precision of atypical users with the item-based technique.

Errors Associated with Atypical Users in the Matrix Factorization Technique.
In this section, we seek to study how the identification measures behave when using a matrix factorization-based technique. We will compare their accuracy to the item-based and user-based. The errors associated with $CorrKMax$ are not studied here, as $CorrKMax$ is dedicated to the memory-based approaches (whether item-based or user-based). Figures 5 and 6 presents the distributions of the RMSE and Precision of the three Abnormality measures with a matrix factorization technique, as well as the reference distribution on the complete set of users.

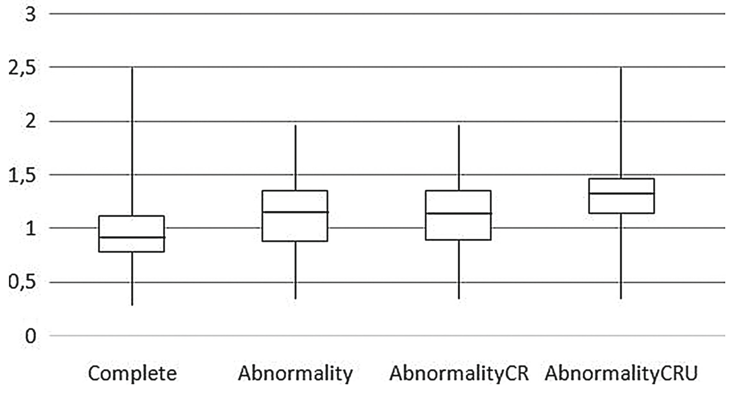

Fig. 5. Distribution of RMSE of atypical users with the matrix factorization technique.

The median RMSE (see Fig. 5)on the complete set of users is equal to 0.92 and the median RMSE obtained with *Abnormality* is 1.17, which corresponds to an increase of 27 %. For the first time, *AbnormalityCR* obtain approximately the same results than *Abnormality*, it has a median RMSE of 1.13. The controversy on resources seems to have no impact on the selection of atypical users with

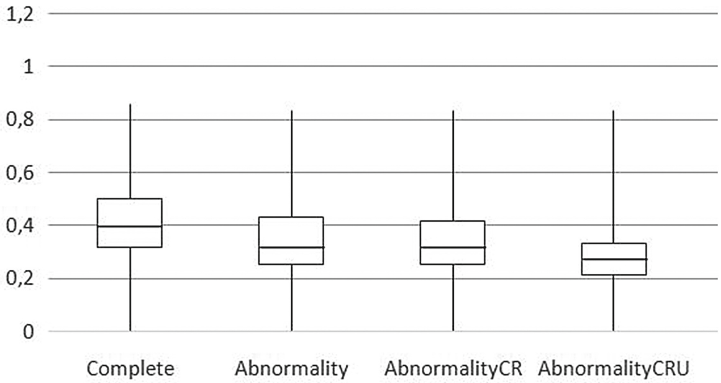

Fig. 6. Distribution of Precision of atypical users with the matrix factorization technique.

the matrix factorization technique. Nevertheless, *AbnormalityCRU* remains the most accurate measure by far for identifying atypical users. Moreover, we can observe that the accuracy of *AbnormalityCRU* is similar to the one observed with the memory-based approaches: 75 % of users identified as atypical belong to the set of 25 % of users who get the worse recommendations in the complete set of users.

In conclusion, we can say that the *AbnormalityCRU* measure, which we propose, is the most accurate measure: when it identifies a user as atypical, he/she most likely will receive low quality recommendations. Moreover, this measure is independent of the recommendation technique: it is efficient on both memory-based (item-based and user-based) and on the matrix factorization model-based approach. [8] has shown that different recommendation approaches (collaborative user-user, collaborative item-item, content, etc.) tend to fail on the same users. It would be interesting to compute the *AbnormalityCRU* on those users.

However, although *AbnormalityCRU* has a high accuracy, some users (from the complete set) with a high per user RMSE of the matrix factorization technique are identified by none of the Abnormality measures: it concerns 50 % of the users who have a RMSE greater than 1.5 (27 users). This means that further work has to be conducted to identify the characteristics of these users.

On the Fig. 6 we observe the same slight difference between *Abnormality* and *AbnormalityCR* than with the user-based technique. This observation should be studied in a future work. The same conclusions can be extracted from those repartitions of Precision, *AbnormalityCRU* is the better indicator of low quality recommendations.

4.5 Synthesis of Results

The correlations between $CorrKMax$ and errors are not conclusive, such as the errors of users selected with this measure. The $CorrKMax$ measure does not allow to identify users who will receive low quality recommendations. This indicator can not be used with any of these recommendation techniques.

At the opposite, the correlations between the abnormality measures (*Abnormality*, *AbnormalityCR* and *AbnormalityCRU*) and RMSE are significant, whatever is the recommendation technique. The *Abnormality* measure from the state of the art allows to select users with a median RMSE/Precision higher than the median RMSE/Precision of the complete set of users, regardless the recommendation technique. This measure shows its best RMSE values on the matrix factorization technique, with an increase of 27 % of the median RMSE, compared to the median RMSE of the complete set. In parallel, with the user-based technique, the *Abnormality* measure obtains the best results with a decrease of 35 % of the median Precision (compared to the complete set).

The *AbnormalityCR* measure shows slightly better results than the *Abnormality* measure except with the matrix factorization technique. Using an item-based technique, the *AbnormalityCR* measure shows its best results with an increase of 29 % of the median RMSE of the complete set of users and an increase of 40 % of the median Precision.

Finally, computing the *AbnormalityCRU* measure remains the best option to be able to identify atypical users, whatever is the recommendation technique. Indeed, *AbnormalityCRU* selects always at least 75 % of users which belong to the worst 25 % of RMSE of the system. Moreover, the *AbnormalityCR* measure does not improve the performance of the state of the art measure with all the recommenders, *e.g.* the matrix factorization technique. The controversy of items seems to improve the performance of the detection only with an item-based technique. Since the *AbnormalityCRU* measure is more complex to compute, the *AbnormalityCR* measure can be a good measure to select atypical users with the memory-based techniques (item-based and user-based), and the *Abnormality* measure would be used with a matrix factorization technique.

5 Conclusion and Perspectives

Social recommender systems is the context of this work. Our objective was to identify users who will receive inaccurate recommendations, upstream of the recommendation process, *i.e.* based only on the characteristics of their preferences. We hypothesized that users with preferences that differ from those of other users will receive inaccurate recommendations. We have referred these users to as atypical users. To validate this hypothesis, we proposed several measures for identifying atypical users, based on the similarity of users preferences with other users, on the average discrepancy of the ratings they provide in comparison with the average rating of other users, on the consensus of ratings on resources, or on users rating profile. We have shown, on a state of the art dataset, that the measure that uses all these criteria is the most accurate one and allows to reliably anticipate that a user will get inaccurate recommendations, with either a KNN-based techniques (user-based or item-based) or a matrix factorization technique.

In a future work, we will focus on the proposition of a new recommendation approach, to provide atypical users with high quality recommendations. In parallel, it will be interesting to investigate the reasons why some users do get inaccurate recommendations and are not identified by any of the measures studied, as mentioned in the previous section. Specifically, a user may be atypical on a subset of items, which is not considered by the measures studied here.

References

1. Goldberg, D., Nichols, D., Oki, B., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM **35**(12), 61–70 (1992)
2. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl. -Based Syst. **46**, 109–132 (2013)
3. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW 1994, (New York), pp. 175–186. ACM (1994)
4. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4:2 (2009)
5. Castagnos, S., Brun, A. Boyer, A.: When diversity is needed... but not expected!. In: IMMM, The Third International Conference on Advances in Information Mining and Management (2013)
6. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.: Generative models for cold-start recommendations. In: Proceedings of the 2001 SIGIR workshop on recommender systems (2001)
7. Grcar, M., Mladenic, D., Grobelnik, M.: Data quality issues in collaborative filtering. In: Proceedings of ESWC- 2005 Workshop on End User Aspects of the Semantic Web (2005)
8. Ekstrand, M.: Towards Recommender Engineering. Tools and Experiments for Identifying Recommender Differences. PH.D. thesis, Faculty of the University of Minnesota (2014)
9. Del Prete, L., Capra, L.: Differs: a mobile recommender service. In: Proceedings of the Eleventh International Conference on Mobile Data Management, MDM 2010, (Washington, USA), pp. 21–26, IEEE Computer Society (2010)
10. Ormándi, R., Hegeds, I., Csernai, K., Jelasity, M.: Towards inferring ratings from user behavior in bittorrent communities. In: Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), pp. 217–222 (2010)
11. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
12. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.T.: Application of dimensionality reduction in recommender system - a case study. In: ACM WebKDD Workshop (2000)
13. Billsus, D., Pazzani, M.J.: Learning collaborative information filters. In: Proceedings of the Fifteenth Intternational Conference on Machine Learning, ICML 1998, (San Francisco, CA, USA), pp. 46–54, Morgan Kaufmann Publishers Inc. (1998)

14. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the Eighth IEEE International Conference on Data Mining, ICDM 2008, (Washington, DC, USA)pp. 263–272, IEEE Computer Society (2008)
15. Haydar, C., Roussanaly, A., Boyer, A.: Clustering users to explain recommender systems' performance fluctuation. In: Chen, L., Felfernig, A., Liu, J., Raś, Z.W. (eds.) ISMIS 2012. LNCS, vol. 7661, pp. 357–366. Springer, Heidelberg (2012)
16. Ghazanfar, M., Prugel-Bennett, A.: Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution. In: 2011 International Conference on Information Systems and Computational Intelligence, 18–20 January 2011
17. Bellogín, A., Castells, P., Cantador, I.: Predicting the performance of recommender systems: an information theoretic approach. In: Amati, G., Crestani, F. (eds.) ICTIR 2011. LNCS, vol. 6931, pp. 27–39. Springer, Heidelberg (2011)
18. Griffith, J., O'Riordan, C., Sorensen, H.: Investigations into user rating information and predictive accuracy in a collaborative filtering domain. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC 2012, (New York), pp. 937–942. ACM (2012)
19. Ekstrand, M., Riedl, J.: When recommenders fail: predicting recommender failure for algorithm selection and combination. In: Proceedings of the sixth ACM conference on recommender systems, pp. 233–236. ACM (2012)
20. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
21. Bellogín, A., Said, A., de Vries, A.P.: The magic barrier of recommender systems – no magic, just ratings. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 25–36. Springer, Heidelberg (2014)
22. Hawkins, D.M.: Identification of outliers, vol. 11. Springer, New York (1980)
23. Aggarwal, C.C.: An introduction to outlier analysis. In: Aggarwal, C.C. (ed.) Outlier Analysis, pp. 1–40. Springer, New York (2013)
24. Bobadilla, J., Ortega, F., Hernando, A.: A collaborative filtering similarity measure based on singularities. Inf. Process. Manage. **48**, 204–217 (2012)
25. Schickel-Zuber, Vincent, Faltings, Boi V.: Overcoming incomplete user models in recommendation systems via an ontology. In: Nasraoui, Olfa, Zaïane, Osmar R., Spiliopoulou, Myra, Mobasher, Bamshad, Masand, Brij, Yu, Philip S. (eds.) WebKDD 2005. LNCS (LNAI), vol. 4198, pp. 39–57. Springer, Heidelberg (2006)