Acquiring the Preferences of New Users in Recommender Systems: The Role of Item Controversy

Igor Sampaio and **Geber Ramalho**[1] and **Vincent Corruble**[2] and **Ricardo Prudêncio**[3]

Abstract. When dealing with a new user, not only Recommender Systems (RS) must extract relevant information from the ratings given by this user for some items (e.g., films, CDs), but also it must be able to ask the good questions, i.e. give a minimum list of items which once rated will be the most informative. Previous work proposed the use of item's controversy and popularity as criteria for selecting informative items to be rated. These works intuitively pointed out at possible limitations of controversy measures with respect to the number of ratings. In this paper, we show empirically that the number of ratings is relevant; we propose a new selection measure of item's controversy; and, we demonstrate that this measure naturally also takes into account the popularity criterion. The experiments showed promising results for the new controversy measure when compared to benchmark selection criteria.

1 INTRODUCTION

A Recommender System (RS) suggests to its users new information items (e.g., films, books, CDs) based on knowledge about their preferences [10]. Despite the apparent success of RS, the acquisition of user preferences remains a bottleneck for the practical use of these systems, even to those applying Collaborative Filtering (CF) - the most popular approach. In order to acquire information about the user preferences, RS commonly present a list of items to the user and ask for his/her ratings.

More specifically, RS have difficulties in dealing with a new user, since they have initially no information about the preferences of such user, and they cannot demand too much effort from him [8]. Indeed, answering many questions or rating numerous items may discourage the user to continue using the system. In this context, besides extracting as much relevant information as possible from the answers given by the user, a RS must also ask good questions, i.e. give a minimum list of items which once rated will be the most informative.

Some previous work [8], [9], [11], following the active learning approach, tried to find out some criteria for selecting the most informative items to be rated. Among these criteria, item's controversy and popularity seemed to provide good results. Controversial items have high entropy, in terms of the ratings given to them, and then, can provide discriminating information. Popular items can provide more information than unpopular ones, since they have more ratings available [9].

[1] Centro de Informática - CIn/UFPE - Cx. Postal 7851, 50732-970, Recife, Brazil, email: ias@cin.ufpe.br, glr@cin.ufpe.br
[2] Laboratoire d'Informatique de Paris VI- LIP6 - 4, Place Jussieu, 75232, Paris, France, email: Vincent.Corruble@lip6.fr
[3] Departamento de Ciência da Informação - DCI/UFPE - Av. dos Reitores, s/n - 50670-901, Recife, Brazil, email: prudencio.ricardo@gmail.com

This work intuitively argued that controversy measure of an item's rates could present limitations if the number of ratings for that item was not taken into consideration (e.g., the controversy over an item only based on 2 ratings is probably not as trustworthy as the controversy over an item with 100 ratings). Based on this intuition, a preliminary solution for this potential problem was proposed: to use a fixed number of ratings for all items [9].

In this work, we present an analysis of controversy as an useful criterion for selecting informative items. In particular, we empirically prove that the number of ratings must be considered and we propose a new selection measure of item's controversy. Finally, we demonstrate that the new measure also takes into account the popularity criteria naturally. The proposed measure was implemented in a KNN-based collaborative filtering algorithm, and evaluated on a database of user ratings for movies. The experiments showed promising results when compared to benchmark selection criteria, particularly for startup databases, for which the limitations of the controversy measure would be more drastic.

In the next section, we describe the new user problem in Recommender Systems. In the Section 3, we discuss previous and related research work. In Section 4, we present a new controversy measure that was called Deviation as well as how it is expected to solve the problems inherent to the variance measure. In Section 5, we describe the experiments organization and the achieved results. At the end, the conclusions and future work are presented.

2 THE NEW USER PROBLEM

A typical problem that arises in RS is the difficulty in dealing with new users, since the system has initially no knowledge about their preferences. This is known in literature of RS as *the new user problem* [1], [8], [11]. On many situations it is possible for the RS to present some items for the new user to rate.

Obviously, the system should not present to the user an exhaustive list of items, since the task of providing the opinions will be tedious and inefficient. In this context, the RS should be able to minimize the user's effort by selecting and presenting those items that would be the most informative for performing good recommendations of new items. This approach is based on *Selective Sampling* techniques used at some *Active Learning* algorithms [6]. The following section discusses item selection strategies based on the concepts of controversy and popularity of an item.

3 PREVIOUS WORK

As previously mentioned, a commonly suggested approach for speeding up the acquisition of a new user is to select the items that,

once rated, will provide the RS with the most information about user preferences. Some previous authors have suggested the use of an item's controversy (entropy) and popularity in order to generate the list of the most informative items that will be presented for the newcomer to rate [8], [9].

Most of the selection methodologies presented in these papers have been applied or tested in RS that use KNN-based Collaborative Filtering (CF) as the way to generate recommendation. This work was also developed with special attention to CF system category

3.1 Popularity

The popularity of an item is given by the number of users who have rated it [8], [9]. So, the more users have rated an item, the more popular it is. In KNN-based Collaborative Filtering, which is the most frequently used recommendation algorithm [5], the suggestions are generated with basis on the opinions of users that are found to be similar to the target user (his neighborhood). Furthermore, the similarity between two users is calculated using the ratings given to the items they have rated in common.

In that context, when a user evaluates a popular item, the system becomes able to determine his similarity with a greater number of other people. So, it is reasonable to expect that rating the most popular items first will result in a much greater information gain.

3.2 Controversy

The use of controversy for choosing the most informative item is based on the intuitive idea that one item (e.g. CD, film) that everybody loved or hated will probably not bring much useful information about the preferences of a new user. This results from the fact that a new user is statistically very likely to be of same opinion that the majority of the other users. Conversely, rating an item for which people have expressed a widely varying range of opinions will probably provide the system with more discriminative information [8], [9].

For measuring the controversy of a given item, a straightforward way is to take the variance of the ratings it has been given [9]. The variance is frequently used to measure the dispersion of a distribution, which makes it reasonably suitable to be used as controversy.

Although using the item controversy as a selection method may provide the system with information that is very discriminative of one's preferences, it only holds true in some situations. The problem occurs when an item is said to have great entropy, but has been rated by a relatively small number of users. For example, in a 1 to 5 evaluation scale, an item with only two ratings, a 1 and a 5, has great entropy but it will probably be of little help in finding user neighborhood or generating recommendations [8].

That fact is especially noticeable when the variance is used as the controversy measure, since it normalizes the dispersion by the number of samples. So, it is possible that, for example, an item with only two ratings will produce the same controversy value as one with a hundred ratings. Teixeira et al. named that as the *problem of width versus intensity of the controversy measure* [9]. In their work, the solution adopted was to define a fixed number of ratings (from now on referred as R) that would be used to calculate the variance. On their work it was adopted $R = 100$, although that decision was not based on an empirical analysis.

That approach brings the constraint that an item must have received a minimum number of ratings to be eligible to be selected. Furthermore, it neglects the information that could be provided by the additional ratings. However it was suggested as a first attempt to solve the width versus intensity problem, and it is also pointed out the need for further and more detailed studies towards a better solution.

4 MEASURING CONTROVERSY WITH DEVIATION

In the previous section we have discussed the problem that may occur when the controversy measure does not reflect on the number of ratings used for the calculation. Therefore, our first aim was to find a controversy measure that would be capable of overcoming the dilemma of width and intensity.

As previously mentioned, the variance formula normalizes the dispersion by the number of samples, consequently eliminating the influence of the number of samples used from the final result. In this case, it could be enough to remove the normalization from the original variance formula. The result is a controversy measure called deviation [3], whose formula is as following:

$$c_i = \sum_{u=1}^{n} (r_{u,i} - \bar{r_i})^2 \quad (1)$$

Where $r_{u,i}$ is the rating given by the user u to the item i, $\bar{r_i}$ is the average of the ratings provided to i and n is the number of evaluations received by i.

By analyzing the formula (1), it is not hard to see that, the greater the number of ratings involved in the calculation, the greater the deviation result will be. The deviation removes the need for estimating a minimum number of evaluations used at controversy computation, as proposed by Teixeira et al.

5 EMPIRICAL ANALISYS

As previously mentioned, the solution proposed by Teixeira et al. for the problem of *width versus intensity* neglects either items that don't have enough evaluation and also the additional ratings.

Furthermore, it is reasonable to suppose that such information loss would have a great impact especially on small databases of starting up systems. In order to investigate that supposition, we decided to run the experiments on a downsized version of the *Eachmovie database* [7]. So, the testing was conducted on a set of 300 randomly selected users and a total of 21518 evaluations and with each user having evaluated at least 30 items.

5.1 Evaluation Metrics

For evaluating the accuracy of recommendation in our experiments we have applied two metrics: ROC [4], commonly used for decision support recommendation and Breese [2], suitable for ranked list recommendation. The use of these two metrics is frequently suggested on Recommender Systems literature.

In order to use ROC, we have considered items evaluated by the users with ratings 1, 2 and 3 as being not relevant for him/her and items rated with 4 and 5 as relevant (as was suggested in [4]). To use Breese, the grade 3 was considered as neutral preference and a half-life of 5 was used, as was suggested in [2].

5.2 Experiments and Results

Our experiments were organized similarly to the ones described in [9]. In that case, the system must choose, one at a time, the items that

will be used to build the user profile. The items evaluated by each user were divided in 3 sets of equal size for 3-fold cross validation.

The process is better described by the algorithm below:

```
Input   U[1..3]: user original items
        subsets to be selected
        n: number of items to select
Output  A: prediction accuracy

UserSelectionTest(U[1..3] , n)
1.  For i=1 to 3
2.    Assign SelectionSet S <- {},
3.           TestSet T <- {},
4.           UserEvaluationsSet E <- {}
5.
6.    T <- U[i]    //a given subset of U
7.    S <- the other 2 subsets of U
8.    While |E| < n
9.        E <- SelectItem(S,E)
10.   P <- Predict(T,E)
11.   a[i]<- Accuracy(P,T)
12.
13. Return average accuracy of a[i],
14.        i = 1...3
```

The function *SelectItem(S, E)* selects one item from the selection set *S* that is not contained in the user evaluation set *E* (the user profile) until it reaches n elements. The function *Predict(T, E)* implements a KNN Collaborative Filtering algorithm as proposed in [2], using the item evaluations present in *E* to generate predictions for the items in the set *T*. Function *Accuracy(P, T)* computes the accuracy of predictions *P* for the items in the set *T* using the metrics ROC or Breese.

In a first experiment run, we aimed to investigate: the real impact of the problem of *width versus intensity* introduced in Section 3.2, the solution of fixing the number of ratings used for variance computation (estimation of parameter R) as proposed in [9] and finally the actual effectiveness of the deviation controversy measure. The following implementations of the function *SelectItem(S, E)* were tested:

- **Random selection**: items are randomly selected. It will be used as a baseline for comparing the selection criteria.
- **Variance**: selection is based on the variance calculated over all the ratings one item has received.
- **Variance with fixed** R: selection is based on the variance calculated over exactly R ratings one item has received. The value of the R parameter was fixed in 20, 40 and 60 and not in 100 as originally proposed [9]. That difference is due to the smaller size of the database used.
- **Deviation**: selection is based on the new controversy measure, the deviation introduced in Section 4.

Figure 1 shows the system's average prediction accuracy using these selection methods. Based on these results, it is possible to say more confidently that there is indeed a problem related to the fact of the variance measure not taking the number of evaluations into account, because the variance used with no restrictions presented the worst of all results. Teixeira et al. pointed out to such problem in [9] but with no further investigation of its real impact. The variance solely is not a good selection method, being even worse than the random selection for most user evaluation sets sizes.

Furthermore, the results of the variance with fixed number of evaluation tend to become better as the value of the parameter R is increased. Even so, the better result was achieved at all points by the deviation measure at both metrics ROC and Breese.

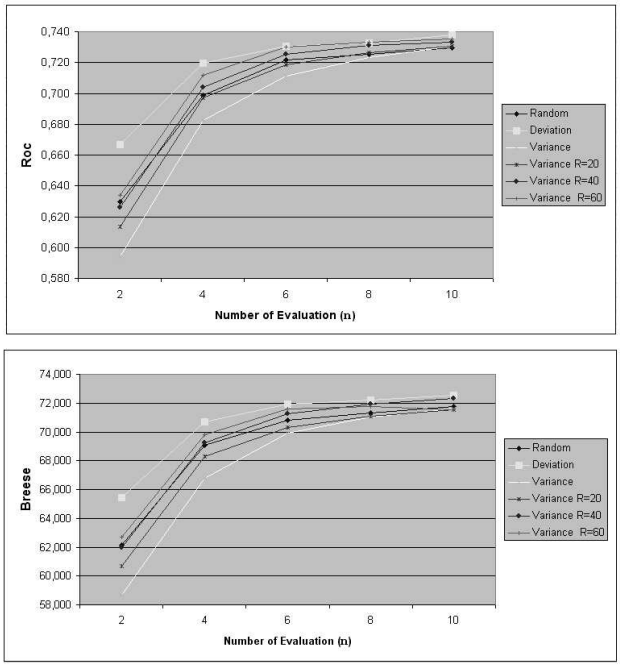

Figure 1. Prediction accuracy using evaluation of items selected by the various controversy criteria and by the random criteria.

An important fact noticed was that, saying that the deviation measure already considers the number of evaluations in its calculation is equivalent to saying that it already takes the **popularity** of the item into account. Indeed, by analyzing the deviation formula, one can realize that is the same as multiplying the variance (controversy) by the number of evaluations (popularity) of an item. This is shown clearly by the formula below:

$$c_i = n \cdot \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_i)^2}{n} \tag{2}$$

So, a second experiment run was made in order to compare the deviation with other selection criteria that combines the concepts of popularity and controversy. This run compared the deviation selection method with the *ActiveCP* selection method [9]. The *ActiveCP* combines the controversy and popularity of an item by generating two item lists. One list containing the items ranked by the controversy with fixed R and the other list with the items ranked by the number of evaluation received (popularity). The two lists are combined to generate a final unique list with the items in the order they should be presented to the user. In the original work the parameter R was set to 100. For the experiments of this work, we decide to use the same 3 values we had previously used, that is 20, 40 and 60.

Figure 2 shows the system's average prediction accuracy using the deviation and the selection method that combines controversy and popularity according to *ActiveCP*.

The obtained results show that, when combined with the popularity, the problem of "width versus intensity" of the variance is greatly diminished. All the combined selection criteria have become very

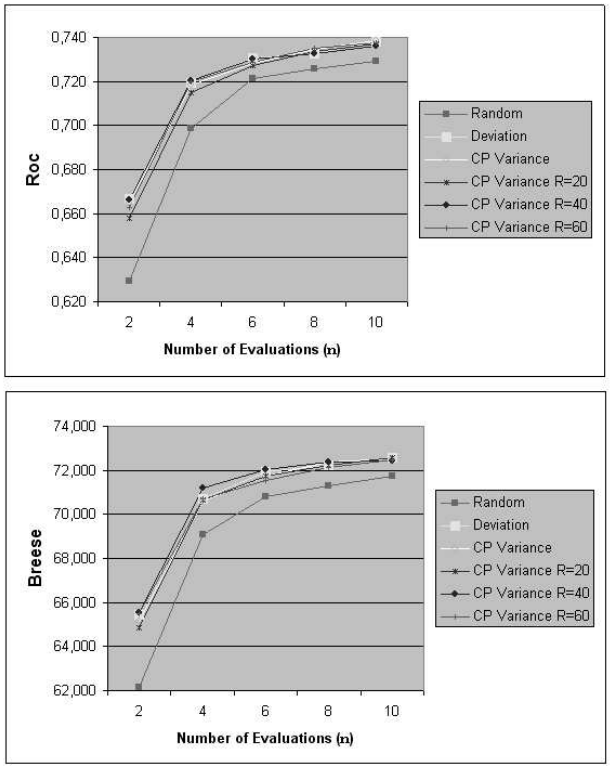

Figure 2. Prediction accuracy using evaluation of items selected by the deviation, *ActiveCP* and the random criteria.

close in prediction accuracy, all significantly better than the random selection. Even so, the deviation has shown to be a rather competitive selection strategy, with results that are best or very close to the best at each profile size. Furthermore, it is not the worst one at any point.

The deviation has some additional advantages. Firstly, there is no need for estimation of the R parameter for finding the value that will provide the better prediction quality. Secondly, it is a method that already combines controversy and popularity in a way that is much simpler than the one proposed in [9] and also computationally faster.

6 CONCLUSIONS AND FUTURE WORK

Previous works have introduced the use of the concepts of controversy and popularity for speeding up the acquisition of new user preferences in Recommender Systems. However, some problem inherently associated with the variance as a controversy measure had been pointed out with no further investigation.

In this paper, we have analyzed the so-called problem of *width versus intensity of the controversy measure*. The impact of the problem was investigated on small, start up like database, in a KNN-based CF Recommender System. The results of the experiments have shown that the problem really exists since the performance was very dependent on the minimum number of evaluations required for variance calculation (i.e., the parameter R). Moreover, in this approach empirical analysis has to be performed in order to estimate an adequate value for the parameter R in each application.

We have then suggested the use of the deviation as a controversy measure capable of solving the limitations of the variance measure. The experiments using the deviation showed that not only does it overcome the problems of the variance measure but it also constitutes a selection method that encompass the concepts of controversy and popularity directly in its calculation. The deviation has the additional advantage that it is much simpler than previous combined methods and also dismisses the need of parameter estimation.

Finally, when using an Active Learning strategy at user preferences acquisition, it is important to consider that a given selected item may not be known by the target user. In that case, the system is causing the user to waste his time with an item he is not able to rate. Measuring how hard it is for the user to sign up in the system (the *user effort* [8]) is also important, but has not been measured by our experiments. Indeed, the experiments described in this work are mainly focused on the prediction accuracy of a system using the discussed selection criteria, but assuming that the user would always be able to rate the items presented to him. Consequently, it would be important to conduct some experiments, for measuring the user effort of the discussed selection criteria.

REFERENCES

[1] C. Boutilier, R. Zemel, and B. Marlin, 'Active collaborative filtering', in *Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 98–106, (2003).

[2] J. S. Breese, D. Heckerman, and C. Kadie, 'Empirical analysis of predictive algorithms for collaborative filtering', in *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 43–52, (1998).

[3] L. Breiman, J.H. Friedman, R. A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Chapman Hall, New York, 1984.

[4] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, 'An algorithmic framework for performing collaborative filtering', in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 230–237. ACM Press, (1999).

[5] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J. Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems*, **22**(1), 5–53, (2004).

[6] M. Lindenbaum, S. Markovitch, and D. Rusakov, 'Selective sampling for nearest neighbor classifiers', in *AAAI/IAAI*, pp. 366–371, (1999).

[7] P. McJones. Eachmovie collaborative filtering data set, http://www.research.digital.com/src/eachmovie (available until october 2004), 1997.

[8] A. Rashid, I. Albert, D. Cosley, S. Lam, S. McNee, J. Konstan, and J. Riedl, 'Getting to know you: Learning new user preferences in recommender systems', in *International Conference on Intelligent User Interface*, pp. 127–134, (2002).

[9] I. R. Teixeira, F. A. T. De Carvalho, G. L. Ramalho, and V. Corruble, 'Active cp: A method for speeding up user preferences acquisition in collaborative filtering systems', in *16th Brazilian Symposium on Artificial Intelligence*, pp. 237–247, (2002).

[10] L. Terveen and W. Hill, *Beyond Recommender Systems: Helping People Help Each Other*, 475–486, Human-Computer Interaction in the Millenium, Addison Wesley, 2001.

[11] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.P Kriegel, 'Probabilistic memory-based collaborative filtering', *IEEE Transactions on Knowledge and Data Engineering*, **16**(1), 56–69, (2004).