# A novel approach to solve the new user cold-start problem in recommender systems using collaborative filtering

Youssouf EL ALLIOUI

**Abstract—** Among the main problems of recommender system one is the "cold start problem". Facing this problem, recommender systems have several methods to overcome the difficulties posed by the initial lack of meaningful data. In this paper, we propose a novel approach to solve new user cold start problem in recommender system applying collaborative filtering. We adopt a mechanism that takes into consideration the new user demographic data and based on similarity techniques finds their 'neighbors'. Our experimental results show the performance of the proposed techniques. We adopt the dataset provided by the GroupLens[1] research team specializing in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems. The proposed system performs better in cases where a large number of users are already registered in the system. In such cases, the system achieves smaller Mean Absolute Error values increasing the accuracy of ratings forecastation. We will provide satisfactory numerical results in different experimental cases.

**Index Terms**— Collaborative filtering, new user cold start, Recommender systems

## 1 INTRODUCTION

Ricci et al. [1] defined the recommender system (RS) as a special type of information system that 1) helps to make choices without sufficient personal experience of the alternatives, 2) suggests products to customers, and 3) provides consumers with information to help them decide which products to purchase. The RS is based on a number of technologies, such as information filtering, classification learning, user modeling and adaptive hypermedia. The recommendation result is the outcome of a complex process that combines the attributes of items and information about users. Recommendation algorithms try, through intelligent techniques, to identify possible connections between items and users and give the most efficient results. The final aim is the maximization of the quality of recommendation (QoR). As QoR could be defined the value of the matching between a specific item and a specific user. In literature, one can find three techniques adopted in RSs:

1) Collaborative filtering (CF) methods : it recommend items to a target user based on given ratings by other users in the community, so it cannot derive an efficient result without the ratings of other users [2] [3] [4] [5] [6] [7] [8] [9] [10].
2) Content-based (CB) methods : it try to match user pro- files against items description and require ratings made by the user herself in contrast [11][12][13][14][15][16].
3) Hybrid methods: have been proposed in order to cover the disadvantages of CF and CB models. These methods combine both techniques in order to provide a more efficient result [10][17].

An important issue for RSs that has greatly captured the attention of researchers is the cold-start problem. Three types of new user cold start problems could be identified: 1) recommendations for new users, 2) recommendations for new items, and 3) recommendations on new items for new users. This problem has two variants: the new user cold-start problem and the new item cold-start problem. The new item cold-start problem occurs when there is a new item that has been transferred to the system. Because it is a new product, it has no user ratings (or the number of ratings is less than a threshold as defined in some equivalent papers) and is therefore ranked at the bottom of the recommended items list. Moreover, this problem can be partially handled by staff members of the system providing prior ratings to the new item. Thus, the concentration of the cold-start problem is dedicated to the new user cold-start problem when no prior rating could be made due to the privacy and security of the system. It is difficult to give the prediction to a specific item for the new user cold-start problem because the basic filtering methods in RSs, such as collaborative filtering and content-based filtering, require the historic rating of this user to calculate the similarities for the determination of the neighborhood. For this reason, the new user cold-start problem can negatively affect the recommender performance due to the inability of the system to produce meaningful recommendations [18]. Addressing this problem has been the primary focus of various studies in recent years.

In this paper, we focus on solving the user side cold start problem. We consider the case where a new user asks for recommendations and no data are available for her

• *Youssouf EL ALLIOUI, LS3M, UHP, BP 145 25000 Khouribga, Morocco. E-mail: youssouf.elallioui@uhp.ac.ma*

[1] https://grouplens.org/

preferences. We propose an algorithm which results the final outcome through three phases. The first phase is responsible to provide means for the classification of the new user in a specific group. For the classification, we adopt efficient techniques like the $C4.5$ algorithm [19] and the Naive Bayes algorithm[20]. In the second phase, the algorithm utilizes an intelligent technique for finding the 'neighbors' of the new user. We examine important characteristics of the user and try to find other users inside the group that best match to her. In the third phase, the final outcome is calculated. This is done adopting prediction techniques for estimating the ratings of the new user.

In comparison with research efforts found in the literature, our approach, 1) handles the new user cold start problem, 2) does not require any a priori probability to be known like efforts adopting probabilistic models, 3) does not require any interview process, 4) does not depend on any complex calculations, 5) involves semantic similarity metrics in the calculation process.

The remainder of the paper is organized as follows. In Section 2, we describe important research efforts in the domain of RSs and the new user cold start problem. Section 3 gives a high level description of the proposed system while. In section 4, we present in detail the key components of our RS. Section 5 is devoted to the presentation of evaluation metrics and the description of our experimental results. Finally, Section 6 concludes our paper.

## 2. STATE OF THE ART

In this section, we briefly present some of the research literature related to recommender systems and the new user cold start problem.

At the milestone of 2017, there are various works aiming to handle this problem. Those studies could be divided into three categories: 1) make use of additional data sources, 2) choose the most prominent groups of analogous users, and 3) enhance the prediction using hybrid methods. The principal idea of the first group is the use of some additional sources, such as the demographic data (a.k.a. the users' profile), the users' opinions, and social tags, for a better selection of the neighbors of the new user [17]. The idea of the second group is to improve the methods that determine the analogous users without the aid of additional data sources [21][22][23]. After determining the most analogous users to the new one, some authors used hybrid methods for the calculation of similarity and/or the prediction of rating [24][25] [25]. This is the basic idea of the third group.

As mentioned, CB systems try to match user profiles against items description. Various techniques have been used in CB models. Vozalis and Margaritis [26] demonstrated a modified version of k-nearest neighborhood by adding a user demographic vector to the user profile and embedding it in the collaborative filtering algorithm for the calculation of similarity. Poirier et al. [27] proposed a method that exploits blog textual data to reduce the cold-start problem by labeling subjective texts according to their expressed opinions to construct a user– item-rating matrix and establishing recommendations through collaborative filtering. Zhang et al. [28] presented a recommendation algorithm that makes use of social tags, particularly user-tag-object tripartite graphs, to provide more personalized recommendations when the assigned tags belong to diverse topics. Almazro et al. [29] introduced a hybrid demographic based and collaborative filtering approach on the movie domain using demographic data to enhance the recommendation suggestion process. Their method classified the genres of movies based on demographic attributes, e.g., user age (child, teenager or adult), student (yes or no), have children (yes or no) and gender (female or male). Preisach et al. [30] argued that many user profiles contain untagged resources that could provide valuable information, especially for the cold-start problem, and proposed a purely graph-based semi supervised relational approach that uses untagged posts. Said et al. [31] [32] modified the user similarity calculation method to employ the hybridization of demographic and collaborative approaches. A modification to the k-nearest neighborhood that calculates the similarity scores between the target user and other users was introduced. Wang et al. [33] introduced Credible and co-clustering filterBot for cold-start recommendations (COBA), which uses the rating confidence level to reduce the dimensionality of the item–user matrix. The items and users were co-clustered, and the ratings within every user cluster were smoothed to overcome data sparsity. The recommendations were fused from item and user clusters to predict user preference. Chen et al. [21] employed additional information, such as the social sub-community and an ontology decision model, to assist the recommendation in the cold-start problem. The social sub-community was divided according to the exiting users' history data and the mining relationship between each other. An ontology decision model was then constructed on the basis of sub-community and users' static information, which makes recommendations for the new user based on his static ontology information. Guo [34] proposed three different approaches from the perspective of preference modeling. First, the ratings of trusted neighbors were merged to form a new rating profile for the active users based on which better recommendations can be generated. Second, a novel Bayesian similarity measure was introduced by taking both the direction and length of rating vectors into account. Third, a new information source called prior ratings, based on virtual product experience in virtual reality environments, was proposed to inherently resolve the concerned problems. Chen et al. [35] proposed a cold start recommendation method for the new user that integrates a user model with trust and distrust networks to identify trustworthy users, which are then aggregated to provide useful recommendations for new users. Demographics data or users' profiles are the most common additional source for solving the cold-start problem. Safoury and Salah [18] presented a framework for evaluating the influence of demographic attributes on the user ratings. This framework was examined using a movie dataset to evaluate the accuracy and precision of the generated recommendations. Formoso et al. [36] proposed a novel profile-expansion approach that includes three types of techniques, namely, item-global, item-local and user-local, based on the query expansion techniques

in information retrieval. The experimental evaluation showed that both item-global and user-local offer outstanding improvements in precision. Son et al. [37] presented a novel filtering method based on fuzzy geographically clustering [38] [39] [40] [41] [42] [43][44], the so-called MIPFGWC-CS, that can handle the issues of selected demographic attributes, the similarities between items and missing ratings that existed in relevant demographic-based algorithms. Finally, Rosli et al. [32] designed a new measure by combining similarity values obtained from a movie "Facebook Page". First, the users' similarities were computed according to the rating cast on the Movie Rating System. Then, the similarity values obtained from a user's genre interest in "Like" information extracted from "Facebook Pages" were combined.

## 3. Methodology

Our approach alleviates the new user cold start problem for RSs applying CF. Fig. 1 present the main operational aspects of this approach.

Let the set of current users in the system be $U = \{u_1, u_2, \dots, u_m\}$, $N = \{n_1, n_2, \dots, n_n\}$ be the set of the new users and $I = \{i_1, i_2, \dots, i_k\}$ an available set of items. There are three phases to predicting item ratings for a new user: 1) User Classification Step, 2) Users Similarity Calculation Step and 3) Users Ratings Forecasted Step.

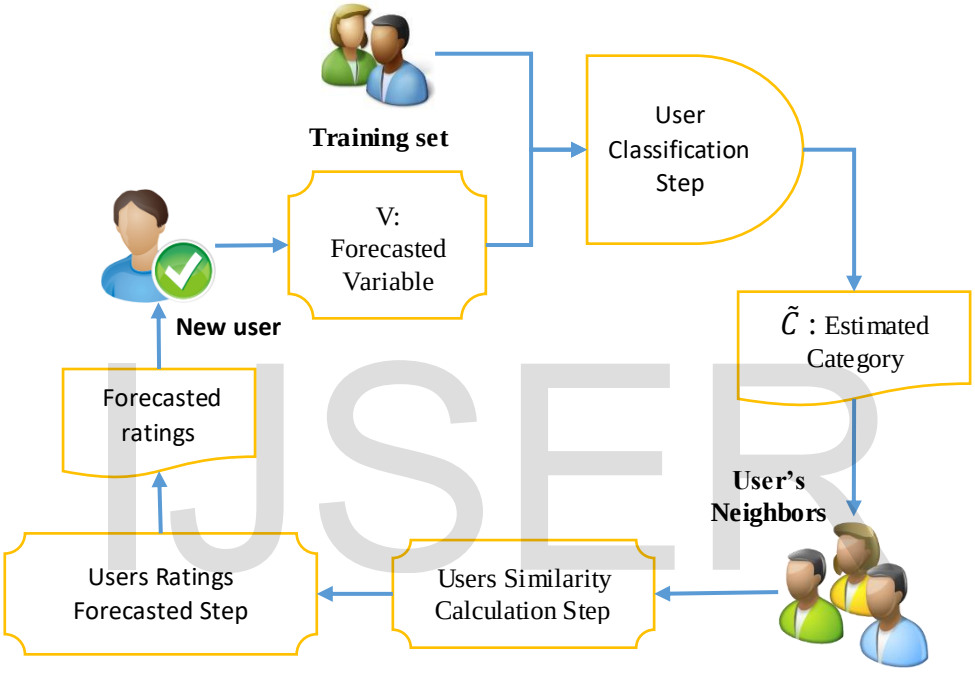


Fig. 1. Proposed approach architecture

### 4.1. User Classification Step

In this step we build a model based on demographic data $D = \{d_1, d_2, \dots, d_l\}$ and users' preferences. Indeed people with a common feedback are much likely to have similar preferences. The classification component implements a model on the basis of a training set that contains instances of the whole data store. Instances include variables related to $D$. Then, we use the generated model to map a new observation in the appropriate category $C$. $C$ belongs in the set of categories $C = \{c_1, c_2, \dots, c_b\}$. In Fig. 1, we show two key factors: $(a)$ the forecasted variable $V$, and $(b)$ the estimated category $\tilde{C}$ . For each new observation $O_j$, $j = 1, 2, \dots$, we set a class attribute that represents $\tilde{V}$. Values of this class are the possible categories $\cup\, c_i \sqsubseteq C$ for each new $O_j$. One of these categories $c_i$ is the output of the model and the corresponding category $\tilde{C}$ for every $n \in N$. The goal is to find a neighborhood $NG = \cup\, u_j\ (NG \in U)\ for\ n \in N$. The neighbors in $NG$ are users that belong to the same category as the model predicts.

Through the use of classifications algorithms, we are able to produce $C$ based on the data related to the set $U$. In order to have the final $C$, we apply binary classifiers while the application of multi class classifiers gives us the opportunity to have multiple classes in the results. In the proposed system, we combine a binary classifier with the model one-vs.-all [46] for achieving multi-class classification. At first, we train the system with the set $U$ and, accordingly, we predict the category $c_j$ for the user $n_j$. In Algorithm 1, we provide the one-vs.-all algorithm.

TABLE 1
ONE-VS.-ALL ALGORITHM

```
INPUT:
   L a learner (training algorithm for
   binary classifiers)
   Samples X
   Labels Y where y_i ∈ {1, …, k} is the label
   for the sample X_i
OUTPUT:
   a list of classifiers f_k, for k ∈ {1, …, k}
```

```
BEGIN
  FOR each k ∈ {1, …, k} DO
    IF y_i = k THEN
      y'_i = 1
    ELSE
      y'_i = 0
    END IF
  END FOR
AND
```

For predicting the new category $c_j$ of a new instance, we apply the generated model. The new category satisfies equation 1:

$$\hat{y} = \arg max_{1 \le k \le K} f_k(x) \quad (1)$$

In this work, we adopt as binary classifier the C4.5 [19] and the Naive Bayes algorithm [20]. In Fig. 1, we see that the result of the discussed process is the estimated category $\tilde{C}$.

### 3.2. Users Similarity Calculation Step

In this step and after the selection of NG, we calculate the similarity between $n \in N$ and each of the neighbors $u_j \in NG$, $j = 1,2,\dots,|NG|$ through a preponderance average of demographic data. We incorporate a similarity function that combines similarity preponderances from different $d_j \in D$, $j = 1,2,\dots,|D|$. In case of numeric data, we use a particular exponential function as described in the following section. For literal attributes, we use a semantic similarity measure [45]. However, in case of binary literal attributes, we take as similarity result boolean values (true or false).

After calculating $\tilde{C}$, we should proceed with grouping users. Our aim is to find the neighborhood of each user $n_j \in N$. The algorithm for finding the neighbors of $n_j$ j is depicted by Algorithm 2. The proposed algorithm aims to match $\tilde{C}_{n_j}$ against $\tilde{C}_{u_j}$, where $\tilde{C}_{n_j}$ is the category of the new user $n_j \in N$ and $\tilde{C}_{u_j}$ is the category of the user $u_j \in U$ The result of the discussed algorithm is the set of neighbors $NG$.

TABLE 2
NEIGHBORHOOD CALCULATION ALGORITHM

```
INPUT: U, N
OUTPUT: NG for each new user
BEGIN
  Define options: one-vs.-all
  Set Class Index
  Build C4.5 Tree
  Build the MultiClass Classifier
  FOR ALL n_i ∈ N DO
    NG = null
    FIND C̃_{n_j}
    FOR ALL u_j ∈ U DO
      IF C̃_{n_j} = C̃_{u_j} THEN
        NG.add(u_j)
      END IF
    AND FOR
  AND FOR
END
```

The next step is to calculate the similarity between new users and users in the $NG$ set. Therefore, the final prediction is based on ratings of the nearest neighbors. The similarity results concern the demographic attributes as defined by $D$. The final similarity degree is calculated through equation 2:

$$sim(n,u) = \frac{\sum_{j=1}^{l} SF_j \,.\, w_j}{\sum_{j=1}^{l} w_j} \quad (2)$$

$SF_j$ is the similarity value of the $j^{th}$ attribute and $w_j$ is the corresponding preponderance. Through this equation, we provide a framework where the developer can focus on specific demographic data. For example, let us consider $D = \{d_1 = age, d_2 = \text{occupation}, d_3 = \text{genderg}\}$. . The discussed set $D$ D can be easily extended. We can focus on age, if we define $w_1 = 0.5$, $w_2 = 0.25$, $w_3 = 0.25$. When $w_j = 1.0$, the calculation process is fully based on the $j^{th}$ attribute.

For each attribute $d_j$, we define a similarity function $SF(at_1, at_2) \in [0,1]$ that gives the results every similarity value $\boldsymbol{SF_j}$. The terms $at_1$ and $at_2$ are the attribute values for a pair of users under consideration. We consider two attribute categories : (a) numeric, (b) literal. For numerical values, we adopt a function $SF : \Re^+ x \Re^+ \longrightarrow [0,1]$ while for literal values, we adopt semantic similarity techniques. In the discussed example, we consider a $SF$ for defining the preponderance of the age $w_a \in [0,1]$ as follows:

$$w_a = \begin{cases} \left(\frac{|diff|}{diff_{max}}\right)^w & |diff| \le diff_{max} \\ 0 & |diff| > diff_{max} \end{cases} \quad (3)$$

Where $diff$ is the difference in age between two users and $diff_{max}$ is a maximum difference (defined by developers). The $w$ argument is a policy factor. If the developer wants to have an increased value of the preponderance $w_a$ even for large $diff$ values, then she will adopt a very small $w$ value (smaller than 1). The opposite stands when $w$ is large.

For literal attribute values, we adopt the known Wu–Palmer semantic similarity metric [45]. Wu–Palmer metric adopts the known Least Common Subsummer (LCS) technique. This technique finds the common node of the two examined issues in the Wordnet[2] taxonomy. Finally, in the case of binary literal attribute (i.e., gender) or binary numerical attribute values, we consider boolean similarity values (true or false). Hence, $SF(at_1, at_2) = 1$ when $at_1 = at_2$ and $SF(at_1, at_2) = 0$ when $at_1 \neq at_2$.

### 3.3. Users Ratings Forecasted Step

We make predictions combining the similarity measure and neighbors ratings. This component implements a function that makes a prediction for an item $i \in I$. The prediction is

[2] http://wordnet.princeton.edu

derived by a preponderance average of each $u_j$ ratings. More specifically, we combine the similarity preponderances calculated in the previous phase with the ratings of neighbors for the possible recommended item.

For each user $n_j \in N$, the model should provide predicted ratings for every item $i_b \in I$. Every predicted rating $\Re_{n_j,i_b} \in \Re^+$ is a preponderance sum of ratings made by the users in $NG$. Therefore, equation 4 holds true:

$$\Re_{n_j,i_b} = \frac{\sum_{u \in NG} sim(n_j,u) \cdot r_{u,i_b}}{\sum_{u \in NG} sim(n_j,u)} \quad (4)$$

Where $r_{u,i_b}$ is the rating of the user $u$ for the item $i_b$.

Based on the above approach, we aim to enhance ratings that are made by users having large similarity degree with every new user. This is as expected as users having in common a lot of characteristics probably they will have similar item preferences.

## 4. Experiments and Result Analysis

In the first, we define certain performance metrics and, then, present our results. Our aim is to quantify the performance of the proposed model concerning the prediction accuracy and compare the results obtained.

### 4.1. Interpretation of RMSE and MAE

Two metrics are used for prediction accuracy. The first one is the Mean Absolute Error (MAE) and the second is the Root Mean Square Error (RMSE).

$$MAE = \frac{1}{k}\sum_{u,i}|P_{u,i} - r_{u,i}| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{k}\sum_{u,i}(P_{u,i} - r_{u,i})^2} \quad (6)$$

$P_{u,i}$ defines the prediction for user $u$ and for item $i$ while $r_{u,i}$ symbolizes the actual rating. Finally, with $k$ we symbolize the number of items under evaluation. Both metrics are widely used to prediction accuracy in evaluating recommender systems.

### 4.2. Experiments

A number of experiments for a specific dataset are running. The dataset is retrieved by the GroupLens[3] research team. GroupLens provides the MovieLens dataset containing one million ratings for 4000 movies defined by 6000 users. From the set of users, we choose a number as the registered users in the system and the rest are considered as new users. We start from 100 registered users and, in different cases, we increase the number till 5000 users. Through this approach, we try to find out how the system behaves for different numbers of registered users. Ratings are between 1 (minimum value) and 5 (maximum value). All ratings are integer values. For each user, we take the identification number and her demographic data $D = \{d_1, d_2, d_3\} = \{\text{age}, \text{occupation}, \text{gender}\}$. Moreover, we consider that $C = \{c_1, c_2, c_3, c_4\} = \{\text{fun}, \text{intellectual}, \text{adventurous}, \text{romantic}\}$. Both lists D and C could be easily extended.

In our experiments, we adopt two classifiers: the C4.5 algorithm and the Naive Bayes (NB) approach. Additionally, we adopt a technique that randomly classifies each user in C. This methodology is named Random Classification Algorithm (RCA). For the C4.5 case, we examine a case where only two classes are used for the classification of each user (depicted by $C^2$4.5 and a case where multiple classes are considered in the classification process (depicted by $C^M$4.5. We compare results taken by the three discussed models (i.e., C4.5, NB and RCA). We examine a number of cases defined by the values of preponderances for each $d_j$. In Table 3, we give a short description of our arguments.

TABLE 3
EXPERIMENTAL ARGUMENTS

| Argument | Values | | | |
|---|---|---|---|---|
| Algorithms | $C^2$4.5 | $C^M$4.5 | NB | RCA |
| $w_j$ | $\sum_{j=1}^{3} w_j = 1 \; for \; w_j \in [0,1]$ | | | |
| $w$ | 0,8 | | | |

### 4.3. Results and discussion

In Table 4, we depict our experimental cases. These cases are defined through the $w_i$ values. Every combination deals with the argument on which the proposed system pays more attention. For instance, in Case 1, the system focuses primarily on the "age" argument in order to issue the required recommendations. Case 4 is more "fair" as all the demographic data are equally considered.

TABLE 4
FOUR EXPERIMENTAL CASES

| Cases | Preponderances | | |
|---|---|---|---|
| | $w_1$ | $w_2$ | $w_3$ |
| First case | 0.6 | 0.3 | 0.1 |
| Second Case | 0.3 | 0.6 | 0.1 |
| Third Case | 0.3 | 0.1 | 0.6 |
| Fourth Case | 0.33 | 0.34 | 0.33 |

Our results for the first case are showing in Fig. 2.. We depict both MAE and RMSE. For both metrics, the $C^2$4.5 5 algorithm exhibits the best performance. As $|U| = 900$, we take $MAE$ approximately equal to 0,8 and RMSE approximately equal to 1.0. As $|U|$ increases, the system has more data to achieve good performance in the classification process as well in exploiting users demographic information. Hence, the error in the prediction becomes smaller. As expected the $RCA$ algorithm performs worse than the rest.

[3] https://grouplens.org

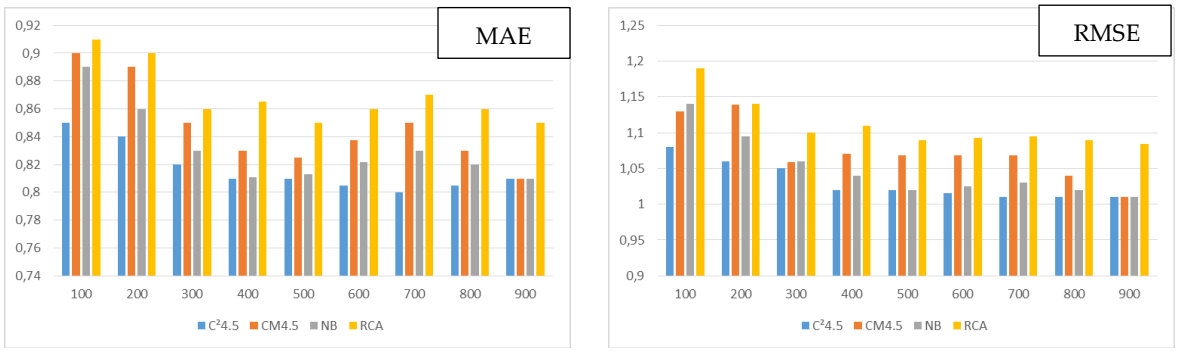


Fig. 2. First case results

In Figs. 3, we see that for the rest of the examined cases, we obtain a similar performance for MAE and RMSE. The $C^2 4.5$ also performs better compared to the rest of the algorithms. Based on these results, we conclude that preponderances for each demographic attribute do not play important role for $|U| \in \{100, 200, \ldots, 900\}$. Hence, we increase the cardinality $|U| \in \{1000, 2000, 4000, 5000\}$.

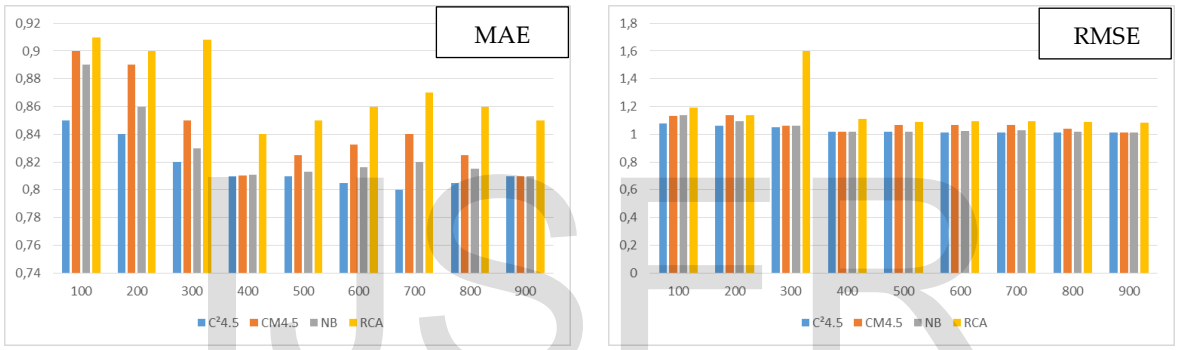


Fig. 3. Third case results

Our results are showing in Fig. 4. Now, the best performance is achieved by the $C^M 4.5$ algorithm accompanied by the $NB$. The minimum MAE value was equal to 0.736 achieved by $C^M 4.5$ when $|U| = 5000$. The minimum value of $NB$ was equal to 0.736 7 for the same number of users. In average, the $C^2 4.5$ 5 algorithm exhibits 0.5% greater MAE value compared to the rest. Through Fig. 4, we see that similar performance is attained for the RMSE metric.

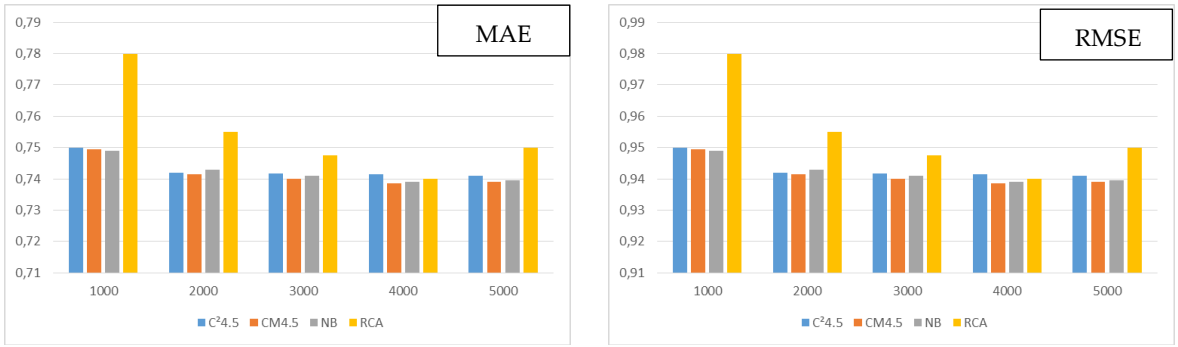


Fig. 4. First case results for multitude users

In this point, we consider $D_{MAE}$ and $D_{RMSE}$ for the MAE and RMSE respectively. $D$ is defined as follows:

$$D = \frac{D_{Base} - D_{Target}}{D_{Base}}\% \quad (7)$$

We try to reveal the difference in the performance when we adopt small number of users ($|U| \in \{100, 200, 400, 500\}$) and when we utilize large number of users ($|U| \in \{1000, 2000, 4000, 5000\}$). $D_{Base}$ stands for $Base \in$

$\{100,200,400,500\}$ and $D_{Target}$ for $Target = 10$.Base. We notice that all the algorithms are affected by the increase in $|U|$. The $C^2$4.5 5 algorithm is less affected compared to the rest. The difference in the performance becomes smaller as $|U|$ increases. However, the difference remains close to 10% as $Base \rightarrow 500$. Concerning the RMSE metric, we see that the $C^M$4.5 algorithm is heavily affected by the increase in $|U|$ as in the MAE case. Smaller $|U|$ leads to greater MAE and RMSE results. This is because the system does not have enough information about users in order to derive better predictions.

Finally, in Figs. 5, we depict our results for $D_{MAE}$ and $D_{RMSE}$ metrics comparing cases where different $w$ values are adopted. We remind that the $w$ argument is a policy factor affecting the preponderance of a demographic attribute (taking numeric values). For these results, we take as Base the case where $w = 15$ and as $Target$ the case where $w = 0{,}8$ In the discussed figure, we see that an increased $w$ value leads to increased $MAE$ values. Better performance is exhibited by the $NB$ algorithm as it is less affected by the change in the $w$ value. When $|U| = 5000$ we observe that the system performs better when $w = 15$. Concerning the RMSE, we see in Fig. 5 that the discussed algorithms exhibit similar behaviour as in the MAE results. Large $w$ values lead to increased RMSE values.

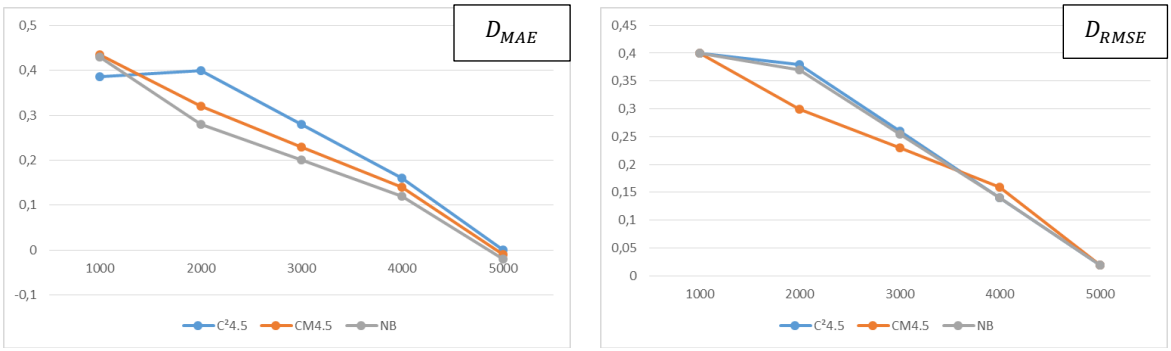


Fig. 5. Results for different $w$ values $D_{MAE}$ and $D_{RMSE}$

## 5. Conclusion

In this paper, we present a novel approach to solve the new user cold-start problem in recommender systems. The idea is that people with a common background and similar characteristics have more possibilities to have similar preferences. Hence, each novel users is classified in a group and accordingly a rating prediction mechanism is responsible to result ratings for items. Our approach is valid for all recommender systems applying the collaborative filtering. It adopts a three-phase to predict item ratings for a new user: 1) User Classification Step, 2) Users Similarity Calculation Step and 3) Users Ratings Forecasted Step. This approach is based on similarity techniques which find the user's neighbors and take into consideration the user's demographic data. Throughout this paper, we use a dataset retrieved by the GroupLens research team specializing in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems. A number of executed experiments show the performance of the proposed techniques. In the case where a large number of users are already registered in the system, smaller MAE values increasing the ratings forecast accuracy are achieved in that system.

## References

[1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to Recommender Systems Handbook," in *Recommender Systems Handbook*, 2011, pp. 1–35.

[2] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme, "Tag-aware recommender systems by fusion of collaborative filtering algorithms," *Proc. 2008 ACM Symp. Appl. Comput. - SAC '08*, p. 1995, 2008.

[3] A. Das *et al.*, "Google news personalization: scalable online collaborative filtering," *Proc. 16th Int. Conf.*, pp. 271–280, 2007.

[4] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the tenth international conference on World Wide Web - WWW '01*, 2001, pp. 285–295.

[5] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*, vol. 4321, 2007, pp. 291–324.

[6] T. Jambor and J. Wang, "Optimizing multiple objectives in collaborative filtering," *Proc. fourth ACM Conf. Recomm. Syst. - RecSys '10*, p. 55, 2010.

[7] R. Jin, L. Si, and C. Zhai, "A study of mixture models for collaborative filtering," *Inf. Retr. Boston.*, vol. 9, no. 3, pp. 357–382, 2006.

[8] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, 2006, p. 501.

[9] M. Khabbaz, L. V. S. L. V. S. Lakshmanan, G. Terms, and S. Descriptors, "Toprecs: Top-k algorithms for item-based collaborative filtering," *ACM Conf. Inf. Knowl. Mangament*, pp. 213–224, 2011.

[10] A. Popescul, D. M. Pennock, and S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," *Proc. Seventeenth Conf. Uncertain. Artif. Intell.*, vol. 2001, no. Uai, pp. 437–444, 2001.

[11] M. J. Pazzani and D. Billsus, "Content-Based Recommendation Systems," *Adapt. Web*, vol. 4321, pp. 325–341, 2007.

[12] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based

Recommender Systems: State of the Art and Trends," in *Recommender Systems Handbook*, 2011, pp. 73–105.

[13] M. Degemmis, P. Lops, and G. Semeraro, "A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation," *User Model. User-adapt. Interact.*, vol. 17, no. 3, pp. 217–255, 2007.

[14] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 54–88, 2004.

[15] D. Billsus and M. J. Pazzani, "User modeling for adaptive news access," *User Model. User-Adapted Interact.*, vol. 10, no. 2–3, pp. 147–180, 2000.

[16] R. J. Mooney and L. Roy and R. J. M. and L. Roy, "Content-based book recommendation using learning for text categorization," *Proc. fifth ACM Conf. Digit. Libr.*, no. June, pp. 195–204, 1999.

[17] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, 2002, p. 253.

[18] L. Safoury and A. Salah, "Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System," *Lect. Notes Softw. Eng.*, vol. 1, no. 3, pp. 303–307, 2013.

[19] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[20] H. Zhang, "The Optimality of Naive Bayes," *Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004*, vol. 1, no. 2, pp. 1–6, 2004.

[21] M. Chen, C. Yang, J. Chen, and P. Yi, "A Method to Solve Cold-Start Problem in Recommendation System based on Social Network Sub-community and Ontology Decision Model," *3rd Int. Conf. Multimed. Technol. 2013) A*, pp. 159–166, 2013.

[22] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Inf. Sci. (Ny).*, vol. 178, no. 1, pp. 37–51, 2008.

[23] M. Daoud, S. K. Naqvi, and T. Siddqi, "An Item-Oriented Algorithm on Cold-start Problem in Recommendation System," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 116, no. 11, pp. 19–24, 2015.

[24] H. Luo, C. Niu, R. Shen, and C. Ullrich, "A collaborative filtering framework based on both local user similarity and global user similarity," in *Machine Learning*, 2008, vol. 72, no. 3, pp. 231–245.

[25] C. W. ki Leung, S. C. fai Chan, and F. lai Chung, "An empirical study of a cross-level association rule mining approach to cold-start recommendations," *Knowledge-Based Syst.*, vol. 21, no. 7, pp. 515–529, 2008.

[26] M. Vozalis and K. Margaritis, "Collaborative filtering enhanced by demographic correlation," *AI Symp. Prof. Pract. AI, 18th World Comput. Congr.*, pp. 1–10, 2004.

[27] D. Poirier, F. Fessant, and I. Tellier, "Reducing the cold-start problem in content recommendation through opinion classification," in *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, 2010, vol. 1, pp. 204–207.

[28] T. Z. Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang, "Solving the cold-start problem in recommender systems with social tags," *a Lett. J. Explor. Front. Phys.*, vol. 92, no. EPL, 2010.

[29] D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, and W. Nzoukou, "A Survey Paper on Recommender Systems," *Arxiv Prepr. arXiv*, vol. abs/1006.5, no. 5, pp. 129–151, 2010.

[30] C. Preisach, L. B. Marinho, and L. Schmidt-Thieme, "Semi-supervised tag recommendation - Using untagged resources to mitigate cold-start problems," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6118 LNAI, no. PART 1, pp. 348–357.

[31] A. Said, E. W. De Luca, B. Kille, B. J. Jain, I. Micus, and S. Albayrak, "KMulE : A Framework for User-Based Comparison of Recommender Algorithms," *Proc. Int. Conf. Intell. User Interfaces*, pp. 323–324, 2012.

[32] A. Said, T. Plumbaum, E. W. De Luca, and S. Albayrak, "A comparison of how demographic data affects recommendation," *User Model. Adapt. Pers.*, p. 7, 2011.

[33] W. Wang, D. Zhang, and J. Zhou, "COBA: A credible and co-clustering filterbot for cold-start recommendations," in *Advances in Intelligent and Soft Computing*, 2011, vol. 124, pp. 467–476.

[34] G. Guo, "Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems," in *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, 2013, pp. 451–454.

[35] C. C. Chen, Y. H. Wan, M. C. Chung, and Y. C. Sun, "An effective recommendation method for cold start new users using trust and distrust networks," *Inf. Sci. (Ny).*, vol. 224, pp. 19–36, 2013.

[36] V. Formoso, D. Fernández, F. Cacheda, and V. Carneiro, "Using profile expansion techniques to alleviate the new user problem," *Inf. Process. Manag.*, vol. 49, no. 3, pp. 659–672, 2013.

[37] L. H. Son, K. M. Cuong, N. T. H. Minh, and N. Van Canh, "An application of fuzzy geographically clustering for solving the cold-start problem in recommender systems," in *2013 International Conference on Soft Computing and Pattern Recognition, SoCPaR 2013*, 2003, pp. 44–49.

[38] L. H. Son, B. C. Cuong, P. L. Lanzi, and N. T. Thong, "A novel intuitionistic fuzzy clustering method for geo-demographic analysis," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9848–9859, 2012.

[39] L. H. Son, B. C. Cuong, and H. V. Long, "Spatial interaction - Modification model and applications to geo-demographic analysis," *Knowledge-Based Syst.*, vol. 49, pp. 152–170, 2013.

[40] L. H. Son, N. D. Linh, and H. V. Long, "A lossless DEM compression for fast retrieval method using fuzzy clustering and MANFIS neural network," *Eng. Appl. Artif. Intell.*, vol. 29, pp. 33–42, 2014.

[41] L. H. Son, "Enhancing clustering quality of geo-demographic analysis using context fuzzy clustering type-2 and particle swarm optimization," *Appl. Soft Comput. J.*, vol. 22, pp. 566–584, 2014.

[42] L. H. Son, "HU-FCF: A hybrid user-based fuzzy collaborative filtering method in Recommender Systems," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6861–6870, 2014.

[43] L. H. Son, "Optimizing Municipal Solid Waste collection using Chaotic Particle Swarm Optimization in GIS based environments: A case study at Danang city, Vietnam," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8062–8074, 2014.

[44] L. H. Son, "DPFCM: A novel distributed picture fuzzy clustering method on picture fuzzy sets," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 51–66, 2015.

[45] M. Palmer and Z. Wu, "Verb semantics for English-Chinese translation," *Mach. Transl.*, vol. 10, no. 1–2, pp. 59–92, 1995.

[46] J. Milgram, M. Cheriet, and R. Sabourin, "'One Against One' or 'One Against All': Which One is Better for Handwriting

Recognition with SVMs?," *Tenth Int. Work. Front. Handwrit. Recognit.*, pp. 1–6, 2006.