Improving New User Recommendations with Rule-based Induction on Cold User Data

An-Te Nguyen
University of Natural Sciences
227 Nguyen Van Cu
District 5, Ho Chi Min City, Vietnam
(84.8) 8353193
nate@hcmuns.edu.vn

Nathalie Denos
Laboratoire LIG
681 rue de la passerelle, BP 72
38402 Saint Martin d'Hères, France
33 (0)4 76 51 49 15
Nathalie.Denos@imag.fr

Catherine Berrut
Laboratoire LIG
681 rue de la passerelle, BP 72
38402 Saint Martin d'Hères, France
33 (0)4 76 51 42 63
Catherine.Berrut@imag.fr

ABSTRACT
With recommender systems, users receive items recommended on the basis of their profile. New users experience the cold start problem: as their profile is very poor, the system performs very poorly. In this paper, classical new user cold start techniques are improved by exploiting the cold user data, i.e. the user data that is readily available (e.g. age, occupation, location, etc.), in order to automatically associate the new user with a better first profile. Relying on the existing α-community spaces model, a rule-based induction process is used and a recommendation process based on the "level of agreement" principle is defined. The experiments show that the quality of recommendations compares to that obtained after a classical new user technique, while the new user effort is smaller as no initial ratings are asked.

Categories and Subject Descriptors
H.3 [**Information Storage And Retrieval**]: H.3.3 Information Search and Retrieval – *information filtering*; H.3.4 Systems and Software – *user profiles and alert services.*

General Terms
Algorithms, Design, Experimentation, Human Factors.

Keywords
Recommender Systems, Collaborative Filtering, Cold start Problem, New-user Problem. Rule-based Induction, Cold User Data

1. INTRODUCTION
In collaborative filtering systems, users receive items on the basis of the ratings they have already provided together with the ratings provided by the other users. The other users are generally compared to user U in order to focus on the ratings made by users that are most similar to U in terms of their past ratings. When a new user connects to the system, his/her list of ratings is empty, which makes it difficult for the system to provide recommendations. This problem is one of the well-known cold start problems, called new-user problem [12]. There are various approaches and techniques to overcome this problem, many of which requiring new users to provide enough ratings for the system to reach a good level of recommendation quality. This first task is both difficult and cumbersome.

In the present work, a general approach is defined, that aims at keeping the new user effort down, by exploiting of all available data about him/her. The term "cold user data", or "cold data", is used to denote the new user data that is available at cold start time. It includes all the data that is readily available or that can be effortlessly collected, as for instance demographic data like age, location, occupation, or any other type of data allowing users to be grouped into communities.

The present work builds on the α-community spaces model [19] [18] that defines a general framework to explicitly group similar users into α-communities. An α-community space is a partition of the set of all users according to the similarity factor denoted by α. In a given collaborative filtering setting, the set of available similarity factors $A = \{\alpha_1, \alpha_2, \ldots, \alpha_{|A|}\}$ is identified, allowing for the building of |A| community spaces. In traditional collaborative filtering, there is one single community space, with α reflecting the similarity of ratings.

In the present work, cold data can serve as a basis for similarity factors α, and allows estimating an *a priori* position for new user U in these α-communities. Collaborative filtering is then processed on each of these alternate α-community spaces. The goal is to obtain similar performances to those obtained with traditional cold start techniques, together with a smaller effort required from new users.

In the remainder of this paper, a state of the art on cold start techniques is made; then the proposal is presented as well as the background knowledge useful for understanding; it is then shown how the approach is applied to the case of MovieLens dataset; finally the experiments are described and the results analyzed, before conclusion.

2. RELATED WORK
Cold start situations occur at the beginning of the use of a recommender system: in such cases, the system lacks data to produce appropriate recommendations. There are 3 different types of cold start problems [2]: new system [15], new item [25] [7],

RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.

and new user. The focus here is on new user cold start situations, where the user has not provided any rating yet, which leads to low recommendation performance.

The answers to this problem vary according to the type of filtering. Except from the case where external sources of information are available on users, the user is always asked to bring a contribution. Depending on the nature of the contribution, the user task is more or less difficult and cumbersome, and the results are more or less reliable.

For collaborative filtering [1], new users are generally asked to rate a set of items; the suggested items have no particular connection with the new user's specific interests. This task is simple, but cumbersome. For instance, MovieLens recommender system [17] requires at least 15 ratings before it is able to provide recommendations, but as the set of proposed movies are not focused on the user, he/she may not be able to rate all of them; as a consequence, he/she may have to go through a very long list of movies to achieve the task. Existing work on this topic aims at finding the best items to be shown to new users, regardless of the particular user at stake [24].

For content-based filtering and hybrid filtering [2], new users are generally asked to define their interest on the basis of a list of terms or example items that best describe them [13] [5]. The required effort is important, as new users have to formulate and synthesize their interests under the form of terms, or have to search for relevant example items. The latter task may be automated when external data on users is available to the system; for instance, it is the case for academic researchers who also are the authors of publications [14] [15]. But in the general case, this kind of data is not available and the resulting profiles are likely to be incomplete and noisy.

Finally, another approach consists in associating any new user with a stereotype among a set of predefined ones. The building process of these stereotypes requires learning data. Then the cold start process requires some data about the new user in order to match him/her to one of the stereotypes. This data is generally collected by the system in an interactive way, for instance by asking a series of questions [10], but also in exploiting demographic resources such as the contents of homepages. In [20], Pazzani studies the intrinsic capacity of the demographic approach to recommendation, in taking new users homepages as an input to automatically feed a demographic recommender system; he concludes that results obtained with the demographic approach are not as good as those obtained with collaborative or content-based systems, but are nonetheless reasonably good.

To conclude, these cold start processes all have a cost for new users, as they require an effort from them, either in terms of time spent, or through the achievement of a difficult or cumbersome task.

This paper studies the quality obtained with an approach where no or little effort is asked to new users. Instead of asking for initial ratings, the system will exploit of available and reliable data on users (for instance demographic data such as age, location, occupation, but also potentially any other type of data that allows grouping users into classes). The presented approach is close to the stereotypes approach, but

- it aims at limiting the number and complexity of the questions asked,
- it intends to make no hypothesis on the type of "cold user data" available both for the learning phase and for the prediction phase (as opposed to Krulwich approach [10] which relies on the availability of a relevant database of demographic data dividing the population into demographic clusters),
- it is meant to be combined with other existing new user approaches in order to improve the cost/benefit ratio for new users.

Furthermore, the framework of α-community spaces in which the approach is developed, is motivated by anticipated further use cases as mentioned in [9], and such as:

- an interactive process allowing users to understand to which similarity factor they can credit a given recommendation (explanation), to visualize their position in the space and move in the various neighborhood spaces (awareness and confidence),
- a recommendation process aiming at providing diversity in recommendations through the various recommendation sources according to various similarity factors,
- an adaptive recommendation process accounting for which similarity factors are most useful to a given user in a given situation (context).

3. PROPOSAL

First, the α-community spaces model is briefly presented, introducing the user "position vector". Then the level of agreement recommendation process is described: it is an *ad hoc* approach that allows for recommendation in the context of α-community spaces. Figure 1 gives an overview of the standard recommendation process proposed. For a given user U, his/her communities are computed, which produces a position vector P. Level of agreement recommendation can then be processed in order to produce a list of recommendations for U.

The last subsection shows how an incomplete position vector can be completed via rule-based induction: it is the heart of the present proposal that allows coping with new user situations by exploiting cold user data.

3.1 The α-Community Spaces Model

The α-community spaces model [18] aims at accounting for all available user similarity factors α that allow grouping users in different ways, as these various groupings are expected

1. to reflect various relevance factors,
2. to bring diversity into the recommendation process via collaborative filtering relating to these various factors,
3. to allow for possible subsequent user interaction with the other users present in the system.

In addition, and this is the main point here, this model shall serve as a basis for capturing new users profiles with a limited effort on their behalf.

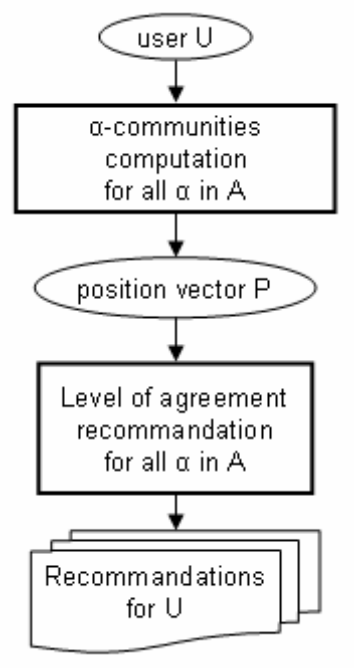

Figure 1. Recommendation process based on α-communities

In a given system setting, the set of available user similarity factors α is identified and denoted by A. For instance in the well-known MovieLens setting [16], age, occupation, location, favorite genre, and ratings are such α ∈ A. Every user U is then associated with a personal position vector P which defines, for each α-space, the community that he/she belongs to (see Figure 2). The picture shows that a user will have different neighbors depending on α. For instance, user U_3 has position vector P_3 = [Academic, Paris, Documentary, Gr#2].

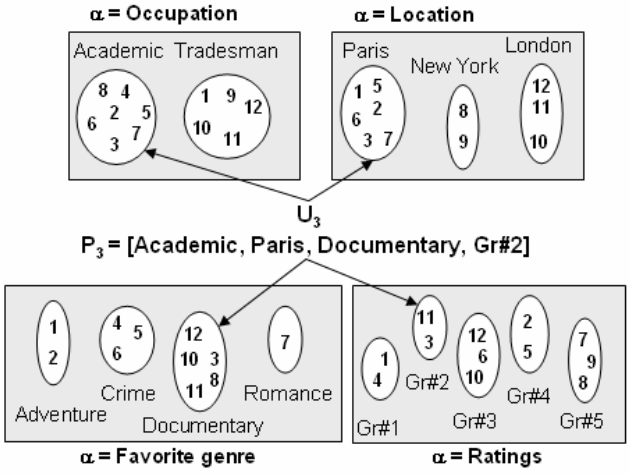

Figure 2. α-communities and example of U_3 position vector

All the position vectors are grouped into a community table (see Table 1) that will be exploited later for induction.

Table 1. Example of an α-community table with |A|=4

Users/ α=	Occupation	Location	Favorite genre	Ratings
U_1	Tradesman	Paris	Adventure	Gr#1
U_2	Academic	Paris	Adventure	Gr#4
U_3	Academic	Paris	Documentary	Gr#2
U_4	Academic	Paris	Crime	Gr#1
U_5	Academic	Paris	Crime	Gr#4
U_6	Academic	Paris	Crime	Gr#3
U_7	Academic	Paris	Romance	Gr#5
U_8	Academic	New York	Documentary	Gr#5
U_9	Tradesman	New York	Documentary	Gr#5
U_{10}	Tradesman	London	Documentary	Gr#3
U_{11}	Tradesman	London	Documentary	Gr#2
U_{12}	Tradesman	London	Documentary	Gr#3

Depending on the nature of α, the way communities are computed will vary. For instance, for α = Age, categories will be defined, such as "children", "teenagers", "adults", "seniors", and users easily categorized with a simple interval-based characteristic function. For α = Favorite genre, as the favorite genre of each user may not be readily available, it has to be computed first on the basis of statistics on user past ratings for instance. For α = ratings, a clustering process will be required to produce α–communities (see [18] for more about this).

To conclude, every user U has a position vector $P = [G_\alpha]_{\alpha \in A}$ where each G_α denotes the α-communities he/she belongs to.

3.2 Ad hoc Level of Agreement Recommendation Process

A plain *ad hoc* recommendation process is proposed: the level of agreement recommendation process. For a given α-community G, which is a set of users that are similar with respect to similarity factor α, a recommendation list is produced. More precisely, this process recommends the items that reach a sufficient "level of agreement" among the members of G, according to the past ratings made by these users.

The selection of recommended items is made as follows: given an α-community G,

1. take the set of movies rated by at least one member of G
2. for each movie compute the average rating among members of G and filter out the movies for which the average rating is lesser than threshold T_{Rating}
3. for each movie, compute the ratio of members of G who rated it (the number of ratings made by members of G, divided by the size of G) and filter out movies for which this ratio is lesser than threshold $T_{Agreement}$

The score of each movie in the recommendation list is the average rating among G members.

For a given user U with position vector $P = [G_\alpha]_{\alpha \in A}$, as many as |A| recommendation lists can be produced with this process: one list per α-community in P.

3.3 Position Vector Completion with Rule-Based Induction

In some cases, some of the positions in the vector P may be missing, or uncertain. In new user situations, positions will be typically missing for α = Favorite genre and α = Ratings, whereas positions will be typically available for α = Occupation or α = Location, as these similarity factors rely on cold user data. For instance, new user U_{13} may have position vector:

$$P_{13} = [\text{Academic, Paris, Crime, \_, \_}]$$

Let A_{Cold} denote the subset of A corresponding to similarity factors α that can be directly obtained from cold user data.

Figure 3 gives an overview of the recommendation process proposed in new user situations. For a given user U, his/her α-communities are computed for $\alpha \in A_{Cold}$, which produces an incomplete position vector P. Rule-based induction is then processed in order to complete the position vector with the help of the past data from the system, i.e. the community table (see Table 1), which contains all the already complete position vectors. Once completed, the position vector allows for level of agreement recommendation on α-communities for all $\alpha \in A$.

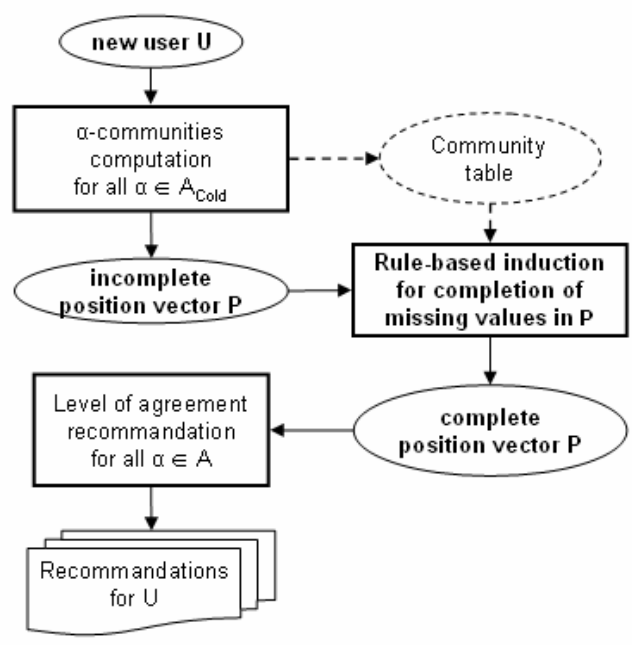

Figure 3. Recommendation process in new user situations

The rule-based induction method [23] [3] [11] [4] induction approach is based on the rough-sets theory proposed by Pawlak [21] in the early 80's. This theory allows to analyze the dependence of a given attribute upon other attributes ; this theory is chosen because of its good performance, but also because:

- it adapts to the symbolic nature of the data,
- it does not require prior knowledge on the data, as opposed to the Dempster-Shafer theory [26] of evidence or to the fuzzy sets theory [29],
- it is based on an extension of set theory, that adapts particularly well to the present formalization of community spaces,
- it offers explanation capacities on communities which may prove useful for subsequent interaction with users
- it can be combined with other approaches, as shown in a number of studies [22], [27], [28].

This type of method begins with a *learning phase* that exploits of the available data to build a classifier under the shape of a set of rules X → d, where X is a subset of A and d is a distinguished $\alpha \in A \setminus X$ called "decision" for which the missing value is to be estimated. For a given user U, X is the set of α corresponding to known values in U's position vector, and d is one of the remaining α for which the value is unknown in P. At cold-start time, X will be generally taken as equal to A_{Cold}.

Then follows the *induction phase* where for a given user U with premise X, an applicable rule r: X → d is searched (see [23] for details).

- If a single applicable rule r is found, the d value replaces the corresponding missing value in position vector P.
- If no rule is applicable, a default value is set for d.
- If more than one rule is applicable, the rule with the higher quality score ρ(r) is chosen.

The quality of a rule ρ(r) is generally based on the two following ratios computed with the community table:

1. the support of r:

$$s(r) = \frac{number\ of\ occurences(r)}{size\ of\ the\ learning\ set}$$

2. the confidence of r:

$$c(r) = \frac{number\ of\ occurrences(r)}{number\ of\ occurrences(X)}$$

The following rule quality measure is adopted:

$$\rho(r) = s(r) \times c(r)$$

In order to measure the quality of the vector completion obtained by this method, level of agreement recommendation will be processed on the corresponding α-communities, and the quality of recommendation will be evaluated.

4. APPLICATION TO MOVIELENS DATASET

This recommendation method is illustrated as applied to the case of the MovieLens dataset in order to show how it contributes to solve the new user problem with the help of cold user data.

4.1 Community Table Computation

The MovieLens dataset [16] allows building 5 community spaces: Age, Occupation, Location, Favorite genre and Ratings. Hence, the community table, that contains all the position vectors for existing users, has 5 columns: age, occupation, location, favorite genre and ratings.

For α = Age, users are grouped into 5 communities on the basis of the following 5 age segments: under 16, 16-25, 26-45, 46-60, over 60.

For α = Occupation, occupations are grouped into the following 7 categories: teacher or student or researcher; tradesman; engineer or technician; artist or leisure; health; retired or at home; others.

For α = Location, communities are built by grouping users according to the state they belong; 44 of them are represented in the dataset.

Both for α = Favorite genre and α = Ratings, communities are formed with a two-step clustering process (see [18] for details) applied on vectors described in the next paragraphs. Firstly, an ant algorithm places users in a 2-D space [6] [8], as motivated by further interactive use of the spaces; for α = Favorite genre, the standard Euclidian distance is used, while for α = Ratings the distance used is Pearson correlation. Secondly, a K-means algorithm builds the communities, chosen for its simplicity and efficiency in most cases.

For α = Favorite genre, the vectors on which clustering is applied are 19-dimension vectors, with one dimension for each of the 19 movie genres. The weight w(U,g_i) for genre g_i and user U reflects

1. the $n_{U,i}$ of movies of genre g_i that user U has rated,
2. the average $\mu_{U,i}$ of these rating,
3. the variance $v_{U,i}$ of these ratings

as follows:

$$w(U, g_i) = a.\frac{n_{U,i}}{N_U} + b.\frac{\mu_{U,i}}{5} + c.\frac{v_{U,i}}{5}$$

with N_U the total number of ratings made by U.

The weights are then normalized so that $\sum_{i=1}^{19} w(U, g_i) = 1$.

Parameters are set as follows: a=0.5, b=0.2 and c=0.3, in order to favor the number of ratings, because users tend to rate movies they like, and not to rate movies they do not like.

For α = Ratings, the vectors on which clustering is applied are the traditional collaborative vectors, filled with the user's ratings.

4.2 Production of Recommendations for New Users

In a new user context, A_{Cold} = {Age, Occupation, Location} denotes α factors that can be obtained directly from cold user data, while α = Favorite genre and α = Ratings will be estimated via rule-based induction.

Figure 4 illustrates a typical situation for new user U with position vector P = [G_{Age}, G_{Occ}, G_{Loc}, _, _]: communities are known for Age, Occupation and Location α-spaces, and values are missing for Favorite genre and Ratings α-spaces.

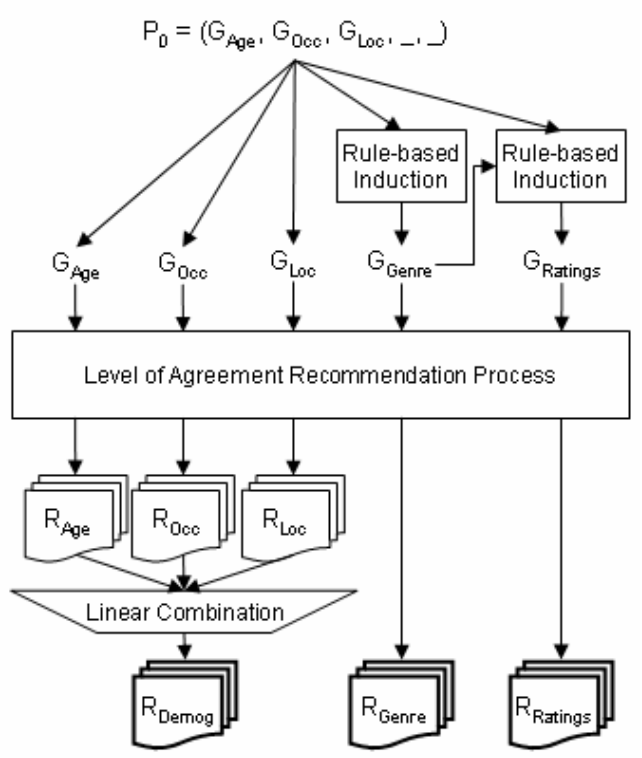

Figure 4. R_i Recommendation lists produced for experiments

Rule-based induction is used to estimate the missing G_{Genre} with X = {Age, Occupation, Location} and the missing $G_{Ratings}$ with X = {Age, Occupation, Location, Genre}.

Then, level of agreement recommendation is processed on each α-space to directly produce 5 recommendation lists: R_{Age}, R_{Occ}, R_{Loc}, R_{Genre}, $R_{Ratings}$, and a 6th list R_{Demog} arising from a linear combination with equal weight of the 3 demographic-like recommendation lists (R_{Age}, R_{Occ} and R_{Loc}).

5. EXPERIMENTS

The aim of the experiments is to see how the recommendation quality obtained with the present method which requires no initial ratings, compares to the quality obtained with a classical Pearson recommendation process [1] after 15 initial ratings. This baseline is reasonable as the level of agreement recommendation process is not an advanced one. The approach will be considered useful if the performance is equal or better than the baseline, as the improvement lies in the reduction of user effort and thus a better cost/benefit ratio.

5.1 Test Data

The MovieLens 100.000 ratings dataset is used, with 943 users and 1682 movies pertaining to 19 genres.

Rule-based induction uses the C5 method [3], which came after the popular C4.5 method [23].

5.2 Protocol and Baseline

Users are divided into two subsets:

1. U_{New} is composed of the 77 users who registered during the last month of April 1998; they play the role of new users
2. U_{Old} is composed of the 866 other users.

For new users, R_{Genre}, $R_{Ratings}$ and R_{Demog} are computed with thresholds T_{Rating} = 4 stars and $T_{Agreement}$ = 25% (see Figure 4).

The baseline is the recommendation list $R_{Pearson}$. It is defined by analogy with the new user technique used by the well-known MovieLens system [17]: a new user is asked for 15 ratings before the first recommendation list is produced. To mimic these conditions, for each user in U_{New} the 15 first ratings are extracted from the dataset, and a recommendation list is produced by classical memory-based collaborative filtering on the basis of these 15-rating initial profiles.

The reference data, to which the R_i lists are compared, is named E; it is the set of ratings assessed by U_{New} users, minus the set of ratings that were used to produce recommendations.

5.3 Evaluation measures

In order to estimate the quality of a recommendation list R produced for a new user U from U_{New}, the precision of prediction is measured with two different metrics.

Firstly, the mean absolute error [24] computes the average of the difference between the scores p_j predicted in R_i and the scores e_j actually assessed by user U in E, with j ranging in the set of items present in the reference set E:

$$\text{MAE} = \frac{1}{|E|}\sum_j |p_j - e_j|$$

As the reference set of ratings E arises from the real use of MovieLens system, there is no guarantee that E contains all the items that users would have (positively) rated if all items had been proposed to them. As a consequence, E may have a low recall.

Secondly, the Pearson correlation can be used to estimate the correlation between the predicted scores and the real ratings:

$$\text{Pearson} = \frac{\sum_j (p_j - \bar{p})(e_j - \bar{e})}{\sqrt{\sum_j (p_j - \bar{p})^2 \sum_j (e_j - \bar{e})^2}},$$

with $\bar{e}$ = average score in E and $\bar{p}$ = average score in R.

5.4 Results

The results are presented in Table 2 for MAE and in Table 3 for Pearson correlation, with the evaluation made for the top 5, top 10, top 15 and top 20 recommendations. The first grey line is the baseline, and the next 3 lines show the average performance for each of the 3 recommendation lists produced as described in Figure 4: R_{Genre} and $R_{Ratings}$ are produced with the help of the proposed rule-based induction; R_{Demog} results from the direct combination of collaborative filtering on the 3 demographic community spaces Age, Occupation and Location.

For good performances, MAE should be low and Pearson correlation should be high. When the proposed approach is as good as, or better than the baseline, the figure is underlined.

Table 2. Results with measure MAE

MAE	Top5	Top10	Top15	Top20	Average
$R_{Pearson}$	1,76%	2,84%	4,09%	5,30%	3,50%
R_{Genre}	1,38%	2,48%	4,09%	5,91%	3,47%
$R_{Ratings}$	3,32%	4,30%	4,95%	4,99%	4,39%
R_{Demog}	*2,59%*	*4,15%*	*4,79%*	*6,51%*	*4,51%*

Table 3. Results with Pearson correlation

Pearson	Top5	Top10	Top15	Top2	Average
$R_{Pearson}$	28,95%	25,90%	24,02%	24,97	25,96%
R_{Genre}	29,92%	30,99%	31,53%	32,23	31,17%
$R_{Ratings}$	31,39%	30,40%	32,13%	30,52	31,11%
R_{Demog}	*18,08%*	*19,45%*	*20,34%*	*19,20*	*19,27%*

R_{Genre} dominates the baseline and all the other recommendation lists, both with MAE and Pearson correlation. R_{Demog} (in italics) is not expected to give good performance, but provides a useful basis for comparison with the 2 other lines. Indeed, the R_{Genre} and the $R_{Ratings}$ recommendations are produced on the basis of the same pieces of information as R_{Demog}, but also benefit from the additional data emerging from the full community table (see Table 1) via rule-based induction.

5.5 Analysis

The experiments show that the approach brings an improvement in new user cold start situations, as the first recommendations made to new users without any initial ratings asked, are as good as, or better than the recommendations made by a classical approach after 15 initial ratings asked.

The two metrics are univocal on R_{Genre}, that is always better than baseline, but not on $R_{Ratings}$, which is worse for MAE and better for Pearson correlation. The fact that $R_{Ratings}$ does not prove as clearly good as R_{Genre}, is probably due to the lesser inductive capacity of the Ratings similarity factor, while Favorite genre is more steady. This phenomenon was studied in previous work [19] where metrics were defined to measure the inductive quality of such factors via an *a priori* analysis of the community table. This work is based on rough sets theory.

R_{Demog} directly exploits of cold user data, whereas R_{Genre} and $R_{Ratings}$ result from a preliminary induction process exploiting of the same data. The results suggest that the induction process brings an improvement.

Nonetheless, there are a number of other factors that may have intervened in the performance of the approach, among which the quality of the methods that compute the community table, the

fine-tuning in the level of agreement recommendation process, and the way demographic recommendation lists are fused. But as no advanced optimization was performed on these underlying processes, the results are even more encouraging with respect to the interest of the approach, as it brings a better cost/benefit ratio for new users in recommender systems.

6. CONCLUSION

To summarize, this paper describes a method that exploits of cold user data to improve the first recommendations provided to a new user, without him/her having to rate any item.

The recommendations are produced with a plain collaborative filtering method applied to various α-community spaces, where α denotes a given user similarity factor. For new users, only some α-communities are known: those relating to cold user data (age, occupation, location, etc.). Plain collaborative filtering on these α-communities leads to bad performance.

The present approach allows estimating the missing α-communities for new users via a rule-based induction process. It is shown that plain collaborative filtering on these estimated α-communities leads to better performance than a classical new user technique. This method could thus be used as a preliminary step before applying another existing new user technique.

The experimental results are encouraging, as they are still positive although the other steps of the recommendation process (production of the initial community table, level of agreement recommendation process, etc.) have not been optimized. Nonetheless further experiments could be done to compare this approach to other more elaborate new user techniques.

This method was applied to the MovieLens context, but it has a more general scope: the α-community space model can be instantiated with other α similarity factors, and the rule-based induction process is also generic for any type of such α.

Moreover, this general setting may also be useful for other situations such as concept drift: when user interests dramatically change, it is often the case that his/her feedback in terms of individual ratings, cannot reflect this change properly or fast enough. The induction process could be used to compute new estimated values for α-communities on the basis of the most successful α factors for this particular user.

7. ACKNOWLEDGMENTS

This work has been partly funded by the French Ministry of Research with the program ACI Masses de Données, project #MD-33.

8. REFERENCES

[1] Breese J. S., Heckerman, D. and Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty In Artificial Intelligence (UAI'98)*, Madison, Wisconsin, USA, 43-52, July 1998.

[2] Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User Adapted Interaction*, *12, 4* (Nov. 2002), 331-370.

[3] C5.0, Release 2.02, September 2005, http://www.rulequest.com/see5-info.html.

[4] CBA v2.1, June 2001, http://www.comp.nus.edu.sg/~dm2/.

[5] Claypool, M,. Gokhale, A. and Miranda, T. Combining Content-Based and Collaborative Filters. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99*, Berkeley, CA, USA, 1999.

[6] Dorigo, M., Bonabeau, E. and Theraulaz, G. Ant algorithms and stigmergy. In *Future Generation Computer Systems (FGCS)*, vol. 16, Elsevier, 851-871, 2000.

[7] Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 439- 446, Orlando, USA, 1999.

[8] Handl, J., Knowles, J., Dorigo, M. On the performance of ant-based clustering. In *Proceedings of the 3rd International Conference on Hybrid Intelligence Systems*, Australia, 2003.

[9] Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22, 1* (Jan. 2004), 5-53.

[10] Krulwich, B., Lifestyle Finder: Intelligent user profiling using large-scale demographic data. *AI Magazine, 18, 2*, 37-45, 1997.

[11] Liu B., Ma Y. and Wong C-K., Improving an Association Rule Based Classifier, In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2000)*, Lyon, France, Sept. 13-16, 2000, 504-509.

[12] Maltz, D. and Ehrlich, E. Pointing the way: Active collaborative filtering. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'95)* (Denver, Colorado, USA). 1995, 202-209.

[13] Melville, P., Mooney, R.J. and Nagarajan, R. Content-Boosted Collaborative Filtering for Improved Recommendations. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI'02*, Edmonton, Canada, 187-192, 2002.

[14] Middleton, S. E., Alani, H., Shadbolt, N. R. and De Roure, D. C. Exploiting Synergy Between Ontologies and Recommender Systems. In *Proceedings of the 11th International World Wide Web Conference WWW-2002, International Workshop on the Semantic Web*, Hawaii, USA, 2002.

[15] Middleton, S.E., Shadbolt, N.R. and De Roure, D.C. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems, 22, 1* (Jan. 2004) 54-88.

[16] MovieLens dataset. http://www.grouplens.org/

[17] MovieLens system. http://www.movielens.umn.edu/

[18] Nguyen, A.-T. *COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés*.

Ph.D. Thesis, University Joseph Fourier, Grenoble, France, 2006.

[19] Nguyen, A.-T., Denos, N. and Berrut, C. Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride, In *Actes de la 3ème COnférence en Recherche Information et Applications (CORIA'06)*, (Lyon, France, March 15-17, 2006), 2006, 303-314.

[20] Pazzani, M., A framework for collaborative, content-based and demographic filtering, *Artificial Intelligence Review*, *13, 5*, 393-408, 1999.

[21] Pawlak Z. Rough Sets. *International Journal of Computer and Information Sciences*, *11 , 5*, 341-356, Plenum Publishing Corporation, 1982.

[22] Pawlak Z., Skowron A. Rough membership functions. In *Advances in the Dempster Shafer Theory of Evidence*, John Wiley & Sons Inc., 251-271, 1994.

[23] Quinlan, J. R., C4.5 : Programs for Machine Learning, *Morgan Kaufmann*, San Mateo, USA, 1993.

[24] Rashid, A., Albert, I., Cosley, D., Lam, S.K., Mcnee, S.M., Konstan, J.A. and Riedl, J. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces, IUI'02*, San Francisco, California, USA, p. 127-134, 2002.

[25] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. 2002. Methods and metrics for cold start recommendations. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 253-260.

[26] Shafer G. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[27] Yao Y.-Y., A comparative study of fuzzy sets and rough sets. *Information Sciences*, *109, 1-4,* 227-242, 1998.

[28] Yao Y.-Y. Generalized rough set models. In *Rough Sets in Knowledge Discovery*, Polkowski L. and Skowron A. (Eds.), Physica-Verlag, Heidelberg, 286-318, 1998.

[29] Zadeh L., Fuzzy Sets, *Information and Control*, *8 (3)*, 1965.