COLLABORATIVE FILTERING ENHANCED BY DEMOGRAPHIC CORRELATION

Manolis Vozalis and Konstantinos G. Margaritis
Parallel Distributed Processing Laboratory
Department of Applied Informatics, University of Macedonia
Egnatia 156, P.O. 1591, 54006, Thessaloniki, Greece
E-mail: {mans,kmarg}@uom.gr

Abstract In this paper we explore how two existing collaborative filtering algorithms can be enhanced by the calculation of demographic correlations among the members of user or item neighborhoods. Experiments are executed to evaluate the performance of the proposed approach. Their results show that demographic data can, in some cases, lead to the generation of more accurate predictions.

Keywords: collaborative filtering, memory-based filtering, demographic data, personalization, prediction, recommender systems

1. Introduction

Recommender Systems were proposed as a computer-based intelligent technique whose purpose was to assist with the problem of information and product overload, by generating suggestions about new items for a particular user. The most common form of input for a recommender system is that of ratings of past items. An alternative kind of input is the demographic data regarding the user or item in mind. This kind of data is usually difficult to obtain and is normally collected explicitly from the user or manually from item catalogues.

Recommender systems are usually distinguished as belonging to one of two wide categories: *Memory-based* or *Model-based Systems*. Memory-based Systems are more efficient as a result of their producing recommendations without a need for any preprocessing. Still, they suffer from serious scalability problems. A different approach is taken by Model-based Systems [2]. These algorithms develop a model of user ratings in order to produce their predictions, which takes time, but once created, speeds up considerably the generation of the recommendations.

Demographic Recommender Systems utilize user attributes, classified as demographic data, in order to produce their recommendations, sometimes with the help of pre-generated demographic clusters. Krulwich [6] and Pazzani [8] have presented systems that rely on demographic data. Hybrid systems [3] combine different filtering techniques in order to produce improved recommendations. Hybrids have been proposed by Balabanovic in Fab [1], Melville in Content-Boosted Collaborative Filtering [7], Claypool in P-Tango [4], GroupLens in filterbots [11], and Smyth and Cotter in their PTV system [12].

In this paper we discuss a novel approach to recommender systems that utilizes two common memory-based algorithms as its base and explores the usefulness of demographic data as an enhancing factor. According to the Burke's taxonomy [3], our algorithm can be classified as a *Feature combination* hybrid, since features from different recommendation data sources are blended into a single recommendation algorithm.

This paper can be outlined as follows. Section 2 provides information about the utilized data set and the evaluation metrics. Section 3 presents an overview of the two filtering algorithms involved in our experiments. Section 4 describes the general structure of the Demographic Algorithm, applies it on the cases of User-based and Item-based Filtering, and concludes with the corresponding experiments. An overall evaluation of the algorithm and pointers for future research are included in Section 5.

2. Experimental Methodology

In order to execute the experiments described in the subsequent sections of this paper we utilized the data publicly available from the GroupLens movie recommender system. The MovieLens data set, used by several researchers, consists of 100.000 ratings which were assigned by 943 users on 1682 movies. Ratings follow the 1(bad)-5(excellent) numerical scale. Starting from the initial data set five distinct splits of training and test data were generated.

Several techniques have been used to evaluate Recommender Systems. We wanted our proposed algorithms to derive a predicted score for already rated items rather than generate a top-N recommendation list. Based on that specific task we selected two evaluation metrics for our experiments: *Mean Absolute Error (MAE)*, a statistical accuracy metric which measures the deviation of predictions, generated by the Recommender System, from the true rating values, as they were specified by the user, and*Coverage*, which measures the percentage of items for which a filtering algorithm can generate predictions.

3. The Base Algorithms

In this section we will briefly discuss the two filtering algorithms which will be utilized by the proposed algorithm: User-based and Item-based Collaborative Filtering. The inspiration for User-based Collaborative Filtering methods comes from the fact that people who agreed in their subjective evaluation of past items are likely to agree again in the future [9]. Similarly to User-based Collaborative Filtering, Item-based Filtering is based on the creation of neighborhoods. Yet, unlike the User-based Collaborative Filtering approach, those neighbors consist of similar items rather than similar users [10].

The execution steps in both cases are similar. First, a suitable data representation is required. That is achieved by the construction of a mxn user-item matrix, R. The next step is Neighborhood Formation, where those users/items that appear to be most similar to the active user/item, i.e. the user/item we wish to generate predictions for, are selected. Those users/items will become the active user/item's neighborhood. The algorithms conclude by Prediction Generation.

4. The Demographic Algorithm

We will present a hybrid algorithm that keeps the core ideas of two existing recommender systems and enhances them with relevant information extracted from demographic data. User-based and Item-based Collaborative Filtering will be our base filtering algorithms. In the approach taken by Pazzani [8], user profiles were expressed as vectors constructed solely from demographic data and similarities among those user profiles were calculated in order for final predictions to be generated. In other words, ratings awarded by the same users for past items were totally disregarded.

In our proposal, user and item correlations, based exclusively on past ratings, lead to the construction of neighborhoods. Still, before these neighborhoods should be utilized for the generation of final predictions, the correlations between neighborhood members and active users or items are re-evaluated, this time by also taking into account existing demographic correlations. Demographic correlations between two users or items, ui_i and ui_j, are defined by the similarity of the vectors which represent the specific users or items. That similarity is calculated by the dot-product of the two vectors. In the following paragraphs we will describe how our general Demographic algorithm can enhance User-based and Item-based Collaborative Filtering.

Table 1. Structure of the User Demographic Vector, *usdemog*

feature#	*feature contents*	*comments*
1	age ≤ 18	
2	$18 <$ age ≤ 29	each user belongs to a single age grouping
3	$29 <$ age ≤ 49	the corresponding slot takes value 1 (true)
4	age > 49	the rest of the features remain 0 (false)
5	male	the slot describing the user gender is 1
6	female	the other slot takes a value of 0
7-27	occupation	a single slot describing the user occupation is 1 the rest of the slots remain 0

4.1 Enhancing User-based Collaborative Filtering with Demographic Correlations

We will first assume that the active user's neighborhood has been already constructed during the Neighborhood Formation step of User-based Collaborative Filtering [13]. In the following paragraphs we will discuss how we can utilize demographic information for prediction generation. In the case of the MovieLens data set, this information includes the age, the gender, a choice from a collection of 21 possible professions, and the zip code for each single user who provided his ratings. Before any of this data could be utilized for the calculation of demographic correlations, via vector similarity, we had to assign a demographic vector for each user in the data set. A user demographic vector, *usdemog*, was defined as a vector with 27 features. Its structure is explained in Table 1.

Once the demographic vectors were constructed for all users, we could now proceed and calculate the proximity between the active user, u_a and the users u_i, for $i = 1, 2, .., l$, belonging to his neighborhood, as it was defined by their registered demographic data. Their demographic correlation, dem_cor_{ai}, was calculated by applying the vector similarity formula on the corresponding demographic vectors. The final step in the recommendation procedure was Prediction Generation. For that purpose, the formula used in plain User-based Filtering [5] was modified as follows:

$$udem\_pr_{aj} = \bar{r_a} + \frac{\sum_{i=1}^{l}(r_{ij} - \bar{r_i}) * enh\_cor_{ai}}{\sum_{i=1}^{l}|enh\_cor_{ai}|}$$

the only difference from prediction generation, as executed in plain User-based Collaborative Filtering, being the enhanced correlation factor. The Enhanced Correlation, enh_cor_{ai}, can be thought as incorporating

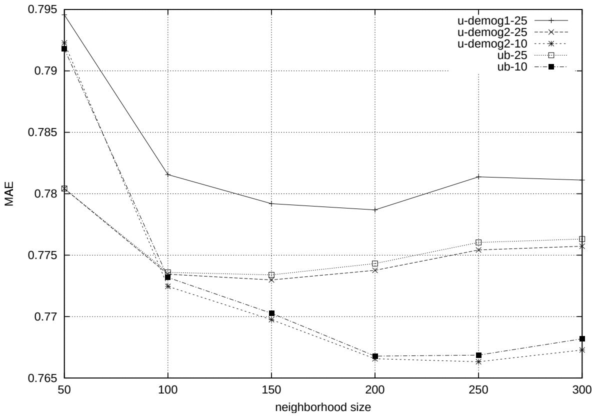

Figure 1. U-Demog1, U-Demog2: Average MAE for various neighborhood sizes

the contributions of the ratings-based correlation and the newly acquired demographic correlation between the active user, u_a, and a neighbor user, u_i. We will now detail two experimental approaches which differ in the way those two correlations are combined.

U-Demog1. For the first set of our experiments we simply *multiplied* the ratings-based correlation, sim_{ai}, and the demographic correlation, dem_cor_{ai}, between the active user, u_a, and a neighbor user, u_i, in order to obtain their enhanced correlation, enh_cor_{ai}. As a result, the Enhanced correlation term was defined as follows:

$$enh\_cor_{ai} = sim_{ai} * dem\_cor_{ai}$$

Figure 1 compares the Mean Absolute Errors (MAE) obtained from U-Demog1 (*u-demog1-25*) and User-based Collaborative Filtering (*ub-25*).

Based on the way that U-Demog1 defines the enhanced correlation factor, we realize that the contribution of demographic correlation in the final prediction is decisive. It is possible for users to be initially included in the neighborhood of the active user, based on their ratings-based correlation with him, and then to see their role in prediction generation diminish, because the value of the demographic correlation disagrees, implying that the same users are not highly correlated with the active user. In the extreme but not improbable case, neighbors who display no common demographic information with the active user will have their involvement in prediction generation completely cancelled out, which will lead to a considerable shrinking of the participating user neighborhood.

The result of possible contradictions between the ratings-based and demographic correlations for the same pair of users is reflected in the results diagram. As we can see, the quality of the predictions generated by U-Demog1 is poorer than those observed for plain User-based Collaborative Filtering. We can conclude that this quality loss is owed to a combination of two factors: a) the shrinking of active user's neighborhood, and b) the inability of demographic correlations to improve on recommendations when playing such a decisive role in prediction generation.

U-Demog2. By this second set of experiments, we decided to let ratings-based correlations be the driving force and assign the demographic component a different role, more as an additive factor on the already set neighborhoods of users, as opposed to *U-Demog1*. Consequently, this time the Enhanced correlation term was defined as follows:

$$enh\_cor_{ai} = sim_{ai} + sim_{ai} * dem\_cor_{ai}$$

Figure 1 compares the Mean Absolute Errors (MAE) obtained from U-Demog2 (*u-demog2-10&25*) and User-based Collaborative Filtering (*ub-10&25*) for two distinct threshold values, *it*={*10, 25*}. The results reported here, were averaged over all 5 data splits.

Based on the way that U-Demog2 defines the enhanced correlation factor, we can realize that demographic correlation can merely *strengthen* the already established, via the ratings-based correlation, participation of a neighborhood user. While there is no change for members of the active user's neighborhood with no demographic resemblance with the active user, a neighbor, u_i, who is highly correlated to the active user, u_a, in terms of demographic information, will see his role in the recommendation procedure increase by a factor of $sim_{ai} * dem\_cor_{ai}$, which grows bigger depending on their calculated demographic correlation.

As we can see from the diagram, the error values of U-Demog2 with it=25 practically coincide with the error values of user-based filtering with the same item threshold. At the same time, the errors of U-demog2 with it=10 are almost identical with the errors of user-based filtering with the same item threshold.

Table 2 includes the best MAE values recorded during the execution of our experiments with U-Demog1 and U-Demog2.

4.2 Enhancing Item-based Collaborative Filtering with Demographic Correlations

Once again, we will assume that the neighborhood formation step from the Item-based Collaborative Filtering procedure has been executed. Before we proceed, we are required to encode the existing demographic

Table 2. Best MAE values of demographically enhanced and plain user-based filtering

	ub-10	*ub-25*	*u-demog1-25*	*u-demog2-10*	*u-demog2-25*
MAE	0,7668	0,7734	0,7792	0,7666	0,7729

information for each item in the data set. The MovieLens data set distinguishes 18 distinct film genres, ranging from Children's to Horror, which results to the corresponding item vector, *itdemog*. It is important to point out that a film can belong to more than one genres at the same time. For example, it can be a Comedy, a Children's flick and a Musical. In that case, the slots which correspond to each of these categories should take a value of 1 (True), with the rest staying fixed at 0 (False). Based on this representation, we proceed with the encoding of all movies from the data set, in each case constructing the suitable demographic vector. The demographic correlation, dem_cor_{jk}, between the active item, i_j, and an item i_k, with $k = 1, 2, ..., l$, which belongs to the neighborhood of the active item, can now be computed by applying the vector similarity formula on the corresponding demographic vectors. The final step in the recommendation procedure is Prediction Generation.

The subsequent experiments test two distinct approaches, whose purpose is to combine demographic correlations with the, already computed, ratings-based item correlations. In both cases, the prediction formula utilized for plain Item-based Filtering [10] is modified to take the following form:

$$idem\_pr_{aj} = \frac{\sum_{k=1}^{l} r_{ak} * enh\_cor_{jk}}{\sum_{k=1}^{l} |enh\_cor_{jk}|}$$

Enhanced Correlation, enh_cor_{jk}, incorporates the contribution of both correlations between the active item, i_j and the neighbor item, i_k.

I-Demog1. In our first set of experiments regarding the utilization of demographic data in Item-based Collaborative Filtering, we defined the enhanced correlation, enh_cor_{jk}, between active item, u_j, and a neighbor item, u_k, by *multiplying* their ratings-based correlation, sim_{jk}, and demographic correlation, dem_cor_{jk}. As a result, the Enhanced correlation term was obtained as follows:

$$enh\_cor_{jk} = sim_{jk} * dem\_cor_{jk}$$

From Figure 2, which compares the Mean Absolute Errors (MAE) collected from I-Demog1 and Item-based Collaborative Filtering (*ib*), we

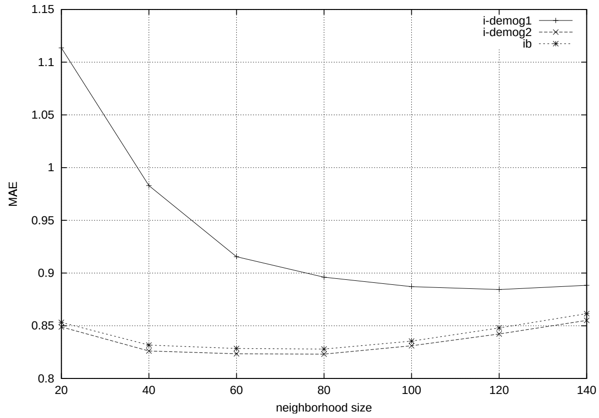

Figure 2. I-Demog1, I-Demog2: Average MAE for various neighborhood sizes

conclude that the recorded results followed the pattern established by the similarly set experiment for U-Demog1. Only this time, the difference in accuracy between the base algorithm, Item-based Filtering, and the demographically enhanced method, I-Demog1, was even wider.

Trying to explain the previous observations, we have to make two important notes.

1 While in the case of user vectors, two individuals have a 50% chance of belonging to the same gender and therefore display a demographic correlation greater than 0, demographic vectors, constructed by the distinct film genres that apply to a specific item, are more difficult to coincide, making it more common for the corresponding items to have a demographic correlation of 0. Based on the selected definition of enhanced correlation, demographic correlations equal to 0 will result to a considerable shrinking of the participating item neighborhood.

2 The experiments for I-Demog1 start with a neighborhood that includes 20 items. This neighborhood size is small and will become even smaller if we take into account all the couples of items with demographic correlation equal to 0. This leads to the poor performance of I-Demog1 for small neighborhoods.

I-Demog2. In this second set of experiments involving the use of demographic data in Item-based Collaborative Filtering, we defined enhanced correlation similarly to U-Demog2:

$$enh\_cor_{jk} = sim_{jk} + sim_{jk} * dem\_cor_{jk}$$

Table 3. Best MAE values of demographically enhanced and plain item-based filtering

	ib	*i-demog1*	*i-demog2*
MAE	0,8279	0,8844	0,8231

Figure 2 compares the Mean Absolute Errors (MAE) obtained from I-Demog2 and Item-based Collaborative Filtering (*ib*).

The results from the figure show that I-Demog2 displays a measurable improvement in accuracy when compared with plain Item-based Collaborative Filtering. Similarly to the U-Demog2 experiments, we realize that demographic data about an item, when merely complementing the ratings-based knowledge we have collected about it, can provide useful information about the item in mind and assist in generating more accurate recommendations. We also have to note that the performance improvement over the base filtering algorithm, documented in the case of I-Demog2, is bigger than the corresponding performance improvement observed in the case of U-Demog2.

Table 3 collects the lowest MAE values which were observed during the execution of the experiments for I-Demog1 and I-Demog2.

5. Conclusions

In this work we have presented a unique filtering approach that draws ideas from existing algorithms and combines them with demographic information available in recommender systems data sets. U-Demog and I-Demog are the hybrid algorithms that resulted by demographically enhancing User-based and Item-based Collaborative Filtering, respectively. We tested two degrees of demographic data involvement in the recommendation procedure. The experimental results showed that our demographically enhanced approaches can vary from being worse than the base filtering algorithms, to outperforming them, depending on the role of the demographic correlations in the prediction generation.

Conclusively, it seems that the demographic information existing in the MovieLens data set do not hold enough data about the users or the items in order to capture their distinguishing features and generate accurate predictions, when utilized just by themselves. Still, when combined appropriately with other forms of filtering, such as collaborative filtering, they can enhance the recommendation process and lead to improved predictions.

References

[1] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40, 1997.

[2] John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, 1998.

[3] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12:331–370, 2002.

[4] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. Combining content-based and collaborative filters in an online newspaper. In *ACM SIGIR Workshop on Recommender Systems-Implementation and Evaluation*, Berkeley, CA, 1999.

[5] Jon Herlocker, Joseph A. Konstan, Al Borchers, and John T. Riedl. An algorithmic frameworkd for performing collaborative filtering. In *The 1999 Conference on Research and Development in Information Retrieval*, 1999.

[6] Bruce Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *Artificial Intelligence Magazine*, 18:37–45, 1997.

[7] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering. In *ACM SIGIR Workshop on Recommender Systems*, New Orleans, LA, 2001.

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

[9] Paul Resnick, Neophytos Iacovou, Mitesh Sushak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, New York, NY, 1994.

[10] Bardul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl. Item-based collaborative filtering recommendation algorithms. In *10th International World Wide Web Conference (WWW10)*, Hong Kong, 2001.

[11] Bardul M. Sarwar, Joseph A. Konstan, Al Borchers, Jon Herlocker, Brad Miller, and John T. Riedl. Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Conference on Computer Supported Cooperative Work*, 1998.

[12] Barry Smyth and Paul Cotter. Surfing the digital wave: Generation personalized tv listings using collaborative, case-based recommendation. In *Third International Conferece on Case-based Reasoning*, Munich, Germany, 1999.

[13] Emmanouil Vozalis and Konstantinos G. Margaritis. Analysis of recommender systems algorithms. In *Proceedings of the Sixth Hellenic-European Conference on Computer Mathematics and its Applications - HERCMA 2003*, 2003.