

# Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System

Laila Safoury and Akram Salah

**Abstract**—Recommender systems have been used tremendously academically and commercially, recommendations generated by these systems aim to offer relevant interesting items to users. Several approaches have been suggested for providing users with recommendations using their rating history, most of these approaches suffer from new user problem (cold-start) which is the initial lack of items ratings. In this paper we suggest utilizing new user demographic data to provide recommendations instead of using rating history to avoid cold-start problem. We present a framework for evaluating the usage of different demographic attributes, such as age, gender, and occupation, for recommendation generation. Experiments are executed using MovieLens dataset to evaluate the performance of the proposed framework.

**Index Terms**—Demographic filtering, information retrieval, personalization, recommender system.

## I. INTRODUCTION

In the recent years, recommender system has been used tremendously academically and commercially providing users with items (i.e.: products, services, or information) which match their preferences and interests. These items are recommended by the system to guide user in a personalized way based on user's historical preferences to discover unseen items among a great collection of items stored on the system. Recommender systems are utilized in different domains to personalize its applications by recommending items, such as books, movies, songs, restaurants, news articles, jokes, among others.

Researchers have suggested several approaches for building recommender systems which offer items differently to users based on a specific assumption in order to match their interests. Nevertheless, all recommendation approaches have strengths and weaknesses that should be considered while choosing the most suitable approach to implement. Therefore, hybrid recommenders are commonly used for combining two or more recommendation approaches together earning better performance and fewer drawbacks [1].

The recommendation systems types can be distinguished into two most commonly used recommendation approaches:

- Collaborative Filtering: Collaborative recommender provides recommendations based on users' similarity, it

assumes that users with similar tastes will rate items similarly. It attempts to find users having similar rating history to the target user (user who requires recommendations), building a neighborhood from which the recommended items are generated.

- Content-based Filtering: Content-based approach recommender offers recommendations based on target user ratings and items associated features, it assumes that user will rate items having alike features similarly. This approach recommends new items having similar features to the items which have been rated by the user.

However, these approaches had been addressed to suffer from new user problem, known as cold-start problem, which is having initial lack of ratings when a new user join the system [2]. Since both approaches assumption are based upon user's ratings history, this problem can significantly affect negatively the recommender performance due to the inability of the system to produce meaningful recommendations [3]. Hence, an alternative kind of input is required to be obtained explicitly from users to be utilized for suggesting recommendations instead of ratings.

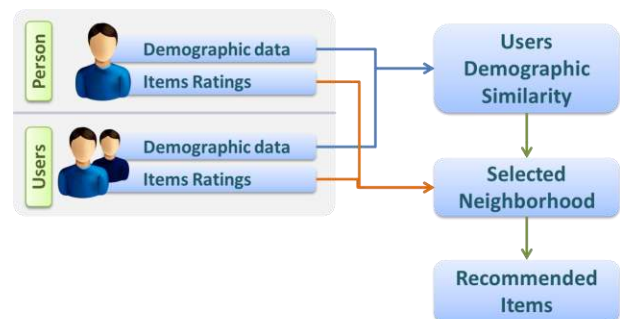


Fig. 1. Demographic-based approach.

Another recommender approach had been introduced which utilizes user demographic data as an alternative input for recommender system which is known as demographic-based approach. Demographic-based recommender, as shown in Fig. 1, suggests utilizing users' demographic data stored on their profiles (i.e. age, gender, location ... etc.), it assumes that users with similar demographic attribute(s) will rate items similarly. This recommender obtains group of user having similar demographic attribute(s) forming a neighborhood from which newly recommended items are generated. In this paper we provide a framework for evaluating users' demographic attributes to be used in generating recommendations for new users.

The rest of the paper is structured as follows. Section II shows other researchers work applying demographic approach in recommender system. In Section III, the

Manuscript received February 9, 2013; revised May 10, 2013.

The authors are with the Computer Science Department, FCI, Cairo University, Cairo, Egypt (e-mail: laila.moustafa@fci-cu.edu.eg, akram.salah@fci-cu.edu.eg).

framework developed in this paper is described. Section IV explains the experiment conducted to evaluate the proposed framework. Finally, Section V concludes the paper and provides directions for future research.

## II. RELATED WORK USING DEMOGRAPHIC RECOMMENDER SYSTEM

The demographic-based and collaborative filtering approaches hybridization had been introduced by researchers for improving the recommendation quality rather than solving “cold-start problem”. A group of researchers have applied a hybrid model-based approach on movie domain using user demographic data to enhance the recommendation suggestion process, it classified the genres of movies based on user demographic attributes, such as user age (kid, teenager or adult), student (yes or no), have children (yes or no) and gender (female or male) [4].

Additionally, other researchers modified user similarity calculation method to employ the hybridization of demographic and collaborative approaches. A modification to k-nearest neighborhood had been introduced which calculates the similarity scores between the target user and other users forming a neighborhood, increasing the scores of users having similar ratings and demographic attribute (each demographic attribute had been evaluated along similar ratings separately) [5]. Whereas another research work demonstrated another modified version of k-nearest neighborhood by adding a user demographic vector to the user profile, the similarity calculation consider both ratings and demographic vector (holding all of the demographic attributes) [6].

In contrast, this paper suggest a novel framework to resolve the new user “cold-start” problem by utilizing the demographic data explicitly given by a user. The framework aims at evaluating the influence of demographic attributes on the user ratings, to assist the recommender system designer to improve recommendations quality for new users. The framework had been examined using a movie dataset to evaluate the generated recommendations accuracy and precision.

## III. DEMOGRAPHIC ATTRIBUTES EVALUATION FRAMEWORK FOR RECOMMENDER SYSTEM

The demographic-based recommendation process performs three stages: data input, similarity calculation and recommendation calculation (as shown in Fig. 2). Data input is the stage which holds new target user’s demographic data (the user who requires recommendations) and also ratings and demographic data of the rest of users. Similarity calculation stage utilizes users’ demographic data to obtain a number of users having similar demographic data to the target user forming a neighborhood. Finally, Recommendation calculation stage obtains items which have been commonly positive-rated by neighborhood users to be suggested to the target user.

Furthermore, the similarity calculation stage requires selecting the demographic attributes to be used for calculating the similarities. For instance, Table I

demonstrates the demographic data of four users; each user has four demographic attributes (gender, occupation, country and age). Let us assume that John is a new user who demand recommendation, the system has to calculate the similarity between John and other users based on the selected attributes. The similarity calculation output depends on the way the system interprets how users are similar, if users having the same occupation are similar then Sarah is similar to John, else if users having the same gender and nationality are similar then Paul and Mike are similar to John. Therefore, the choice of attributes affects the similarity calculation output which consequently influences the results of recommendation calculation stage.

TABLE I: EXAMPLE OF USERS DEMOGRAPHIC DATA

Name	Gender	Occupation	Country	Age
John	M	Student	France	13
Paul	M	Doctor	France	34
Sarah	F	Student	USA	12
Mike	M	Teacher	France	27

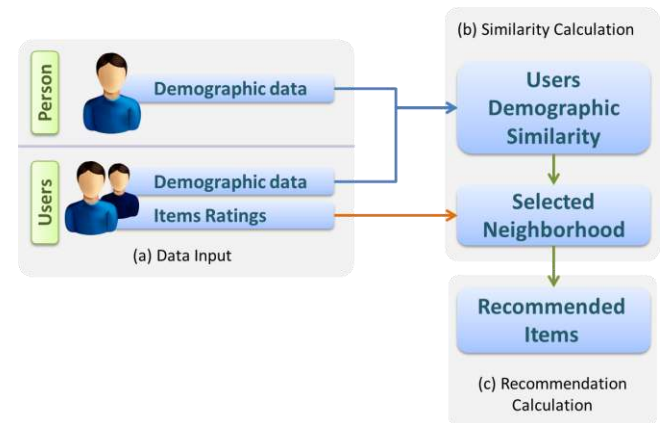


Fig. 2. Demographic-based approach for new users.

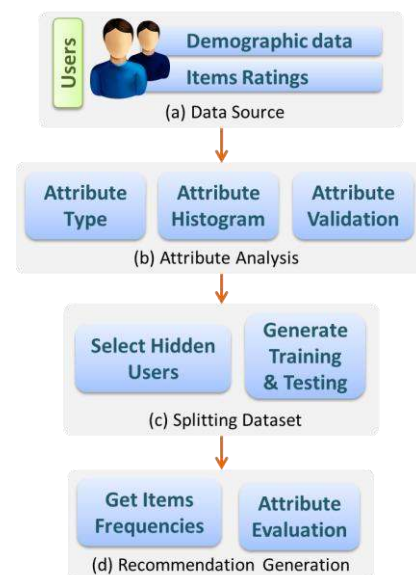


Fig. 3. Demographic attributes evaluation framework.

The proposed framework consists of four modules: data source, attribute analysis, splitting dataset and recommendation generation (as shown in Fig. 3). Data source has all data about users (demographic and items ratings) stored. Attribute analysis module works on analyzing the

type of demographic attributes, the distribution of attributes values across the dataset (histogram) and validity of using these attributes for recommendations. Splitting dataset module splits training and testing for each valid attribute by removing all the ratings of a few randomly selected users (considered as hidden/new users who have no ratings) from training file and adds their ratings to a testing file. Afterwards, Recommendation generation module extracts most frequent items appeared in the training file (rated by users having similar attribute value to the hidden users) recommending them to the new users (hidden users), the testing file is used for evaluating the correctness of the recommendations compared to the hidden ratings.

#### IV. EXPERIMENTAL METHODOLOGY

The framework had been experimented using the publicly available data of GroupLens movie recommender system, MovieLens data set (<http://www.grouplens.org/node/73>). This dataset had been used by many researchers; some researchers used the dataset to execute their experience [6] while others used the MovieLens dataset to study the state-of-art of recommender systems applying collaborative approach different techniques [7]. Additionally, GroupLens provides various versions of the dataset, such as: MovieLens 100k, MovieLens 1M, and MovieLens 10M datasets.

TABLE II: MOVIELENS DATASET INFORMATION

MovieLens Dataset Files	File Attributes Description
u.user	The user file contains demographic information about the 943 users. "user id   age   gender   occupation   zip code"
u.item	The item file holds information about the items (movies). "movie id   movie title   release date   video release date   IMDb URL   unknown   Action   Adventure   Animation   Children's   Comedy   Crime   Documentary   Drama   Fantasy   Film-Noir   Horror   Musical   Mystery   Romance   Sci-Fi   Thriller   War   Western"
u.data	The data file contains 100,000 ratings by 943 users on 1682 items. "user id   item id   rating   timestamp"

##### A. Data Source

The dataset used in this paper is MovieLens 100k; it consists of 100,000 ratings which were evaluated by 943 users on 1682 movies. Each user had rated at least 20 movies; the ratings are assigned numerically from 1(bad) to 5(excellent). Table II shows information about MovieLens dataset files used in the experiment.

##### B. Attribute Analysis

The attribute analysis module determines the type and value ranges of the demographic attributes in MovieLens dataset, shown in Table III. Then, the frequency of each attribute value is calculated showing the number of users having similar value; Fig. 4 illustrates the histogram of MovieLens demographic attributes except for zip code which had only 148 low frequent duplicated values. Afterwards, the module validates the attributes that can be used for

recommendations by checking the following conditions:

- 1) **Invalid Value Range:** It occurs when some ranges of the demographic attribute has low frequency, such as the frequency of age attribute "less of equal nine years old" range has only one user within its range (Fig. 4 (b) ), therefore, the system will not be able to offer recommendations for new users who fall in this range using age attribute.
- 2) **Invalid Attribute Value:** It exists when a value of an attribute has a vague meaning, such as occupation attribute values "none" and "other" (Fig. 4 (c) ), if more than one user had their occupation filled as "none" or "other" it doesn't imply that they are having similar taste or will rate items similarly.
- 3) **Invalid Attribute:** Attribute is considered invalid when its values are highly sparse, such as zip code attribute most of its values are distinct while a few has low frequency.

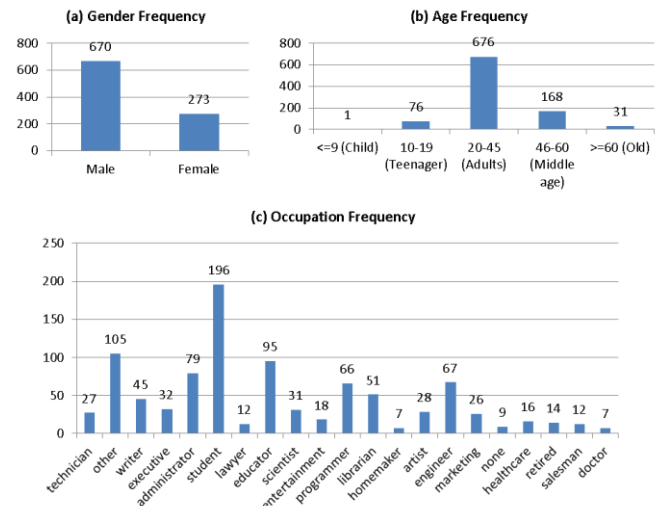


Fig. 4. Attributes Demographic

##### C. Splitting Dataset

TABLE III: ATTRIBUTE TYPES

Attribute Name	Data Type	Value Ranges
Age	Number	7 – 73
Gender	Character	M,F
Occupation	Text	21 Occupations
Zip Code	Text	795 distinct value

TABLE IV: NUMBER OF HIDDEN USERS PER ATTRIBUTE

Attribute	Testing Users Number	Total No.
Age	Teenager ( 10-19) = 10	40 users
	Adults (20-45) = 20	
	Middle age (46-60) = 15	
	Old (>60) = 5	
Gender	20 per 2 genders	40 users
Occupation	10 per 4 occupations	40 users

Splitting dataset module creates training and testing files for each of the three valid attributes (age, gender and occupation) and their valid values (excluding "age" invalid

range and “occupation” two invalid values) to be evaluated. Training dataset requires selecting number of users to hide their ratings adding them to testing dataset. Table IV illustrates the number of users whom their ratings will be removed from training dataset (40 users per attribute); in our experiment only the most four frequent values of occupation attribute: Student, Educator, Administrator, and Engineer, will be considered while the rest of occupations will be excluded for sake of decreasing the number of trials.

#### D. Recommendation Generation

The Recommendation generation module utilizes the training file of each attribute to calculate the frequency of all items rated by users having similar attribute value. For instance, the module uses gender training file to calculate the frequency of items rated by female gender and vice versa for male gender. Table V shows the frequency of a few movies which had been rated by female gender, the columns present the numeric scale rating from 1 to 5 showing how many female gender rated the movie by 1,2,3,4 and 5 while the last column represents the average of movie using equation (1).

$$\text{Average} = ((\text{Freq}_1 * 1) + (\text{Freq}_2 * 2) + (\text{Freq}_3 * 3) + (\text{Freq}_4 * 4) + (\text{Freq}_5 * 5)) / (\text{Freq}_1 + \text{Freq}_2 + \text{Freq}_3 + \text{Freq}_4 + \text{Freq}_5) \quad (1)$$

TABLE V: SAMPLE OF FEMALE GENDER MOVIES FREQUENCIES

Movie id	F <sub>1</sub>	F <sub>2</sub>	F <sub>3</sub>	F <sub>4</sub>	F <sub>5</sub>	Average
1	3	8	24	46	28	4
2	1	1	5	8	2	4
3	3	3	3	3	1	3
4	2	5	12	18	3	3
5	0	3	2	13	3	4

TABLE VI: ACCURACY MEASUREMENTS RESULTS

Attribute	Value	RMSE	MAE	Total RMSE	Total MAE
Gender	Female	0.86	0.88	0.81	0.83
	Male	0.76	0.78		
Occupation	Student	0.85	0.82	0.82	0.81
	Engineer	0.73	0.71		
	Admin	0.83	0.83		
	Educator	0.87	0.88		
Age	Teenager	1.00	1.00	0.88	0.88
	Adult	0.81	0.83		
	Mid-Age	0.79	0.77		
	Old	0.93	0.93		

Now, the module can generate recommendations using the highly frequent positive-rated items for each attribute value offering them to hidden users having equivalent attribute value. In our experiment these suggested items are compared to the hidden users’ ratings, stored in testing dataset, using accuracy prediction measurements which are commonly used by most of recommender systems to evaluate the quality of the generated recommendations. The accuracy prediction measurements used in our work are Mean Absolute Error

(MAE), Root Mean Squared Error (RMSE), Precision, Recall, and F-score [8].

TABLE VII: PRECISION MEASUREMENTS OF TOP 30 MOVIES RATED BY 4

Attribute	Value	Rate 4			Total F-score
		R	P	F	
Gender	Female	19%	18%	16%	14%
	Male	15%	15%	12%	
Occupation	Student	24%	14%	17%	13%
	Engineer	15%	13%	12%	
	Admin	14%	15%	12%	
	Educator	15%	18%	13%	
Age	Teenager	24%	13%	15%	13%
	Adult	16%	14%	13%	
	Mid-Age	14%	19%	13%	
	Old	16%	14%	11%	

TABLE VIII: PRECISION MEASUREMENTS OF TOP 30 MOVIES RATED BY 5

Attribute	Value	Rate 5			Total F-score
		R	P	F	
Gender	Female	22%	16%	16%	17%
	Male	25%	24%	19%	
Occupation	Student	34%	8%	11%	16%
	Engineer	23%	11%	14%	
	Admin	23%	22%	20%	
	Educator	22%	29%	19%	
Age	Teenager	25%	8%	11%	15%
	Adult	21%	22%	19%	
	Mid-Age	27%	25%	19%	
	Old	28%	17%	12%	

Accuracy measurements, MAE and RMSE, evaluate the accuracy of the predicted ratings compared to user actual ratings. We considered the average of movies frequency (Table V) as a predicted ratings which is compared to all hidden users ratings using accuracy measurement. The results of the evaluation are shown in Table VI, it demonstrates that gender and occupation attribute are almost similar whereas age attribute showed further error percentage.

Precision measurements, Precision, Recall and F-score, evaluate the correctness of the recommended item predicted rather than its predicted rating. Precision measures the system’s ability to offer items which is truly relevant for users compared to the total generated recommendations, whereas Recall measures the ability of system to collect relevant items to users compared to users’ actual ratings, and F-score is the harmonic between precision and recall. In our experiment we considered that items which are rated by 4 and

5 are positively relevant for user.

The precision measurements are calculated using hidden users' items rated by 4 and 5 comparing them to the 30 highest frequent items rated by 4 and 5 through other users having similar attribute value to the hidden users. Table VII and Table VIII illustrates precision, recall and f-score of generated recommendations per attribute value for top 30 items rated by 4 and 5 respectively. The total F-score is used to evaluate the whole attribute, it can be implied that the gender attribute has the highest total F-score of its top 30 rated items scored by 4 and 5 whereas other attributes total F-score are almost equal.

## V. CONCLUSION AND FUTURE WORK

In this work we have presented a novel framework for evaluating demographic attributes available in recommender systems datasets to be used for recommending relevant items to new users. The framework was examined using MovieLens dataset, the experimental results of the dataset showed that all attributes have almost the same influence. Conclusively, it seems that the demographic data in the MovieLens dataset does not influence differently on users' ratings.

Further research can be performed to enhance the results, such as creating more than one training and testing dataset to be evaluated and gather the average of the results. Also a higher level of movie recommendation can be obtained by relating the movie genres to demographic attributes. Finally, this framework can be applied on different domains datasets.

## REFERENCES

- [1] R. Burke, "Hybrid web recommender systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, ed., Springer, 2007, pp. 377-408.
- [2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, pp. 734-749, June 2005.
- [3] J. B. Schafer, D. Frankowski, and J. Herlocker, "Collaborative Filtering Recommender Systems," in *The Adaptive Web: Methods and*

*Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, ed., Springer, 2007, pp. 291-324.

- [4] D. Almazro, G. Shahatah, L. Albulkarim, M. Khrees, R. Martinez, and W. Nzoukou. "A Survey Paper on Recommender Systems," *arXiv preprint arXiv:1006.5278*, Dec. 2010.
- [5] A. Said, T. Plumbaum, W. E. De Luca, and S. Albayrak, "A comparison of how demographic data affects recommendation," in *Proc. 19th international conference on User modeling, adaption, and personalization*, 2011.
- [6] M. Vozalis and K. G. Margaritis, "Collaborative filtering enhanced by demographic correlation," in *Proc. AIAI Symposium on Professional Practice in AI, of the 18th World Computer Congress*, 2004.
- [7] L. Candillier, K. Jack, F. Fessant, and F. Meyer. "State-of-the-art recommender systems," in *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*, M. Chevalier, C. Julien, and C. Soule-Dupuy, ed., IGI Global, 2009, pp. 1-22.
- [8] G. Shani and A. Gunawardana, "Evaluating recommendation systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, ed., Springer, 2011, pp. 257-297.



**Laila Safoury** was born in Cairo, Egypt on March 8, 1988. She obtained her Computer Science Bachelor Degree in Faculty of Computers and Information (FCI), Cairo University, Cairo, Egypt, in 2009. She works as a teacher assistant since 2009 at FCI Cairo University in Computer Science Department, Cairo, Egypt. She also worked as a researcher at Center of Excellence in Data Mining and Computer Modeling track "Arabic Documents Sentiment Classification" sponsored ITIDA (2010). She had worked recently as a researcher on "Society in Hand" website project, applying Recommender system approach to the project (2012).



**Akram Salah** was born in Cairo, Egypt and he graduated from Mechanical Engineering in 1973 from Cairo University, Egypt. He got his Ph.D. in computer and information sciences from University of Alabama at Birmingham, USA, 1986. He has taught computer science in University of Alabama at Birmingham, USA, Michigan State University, USA, and North Dakota State University, USA where he started a graduate program in Software Engineering. He also taught in the American University in Cairo, Egypt and he is currently an Associate professor in Faculty of Computer and Information, Cairo University. His research interests are in data engineering, software engineering and knowledge engineering, in which he has over 90 publications. Dr. Akram Salah is a member of ACM and IEEE.