Knowledge-Based Systems 100 (2016) 175–187

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

User profiling approaches for demographic recommender systems

Mohammad Yahya H. Al-Shamri[a,b,*]

[a] Computer Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia
[b] Electrical Engineering Department, Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen

ARTICLE INFO

Article history:
Received 22 September 2015
Revised 4 March 2016
Accepted 5 March 2016
Available online 11 March 2016

Keywords:
Recommender system
User profile
Demographic data
Similarity computation

ABSTRACT

Many of our daily life decisions rely on demographic data, which is a good indicator for closeness of people. However, the lack of these data for many online systems let them search for explicit or implicit alternatives. Among many, collaborative filtering is the alternative solutions especially for e-commerce applications where many users are reluctant to disclose their demographic data. This paper explores, discusses and examines many user-profiling approaches for demographic recommender systems (DRSs). These approaches span many alternatives for profiling users in terms of the attribute types, attribute representations, and the profiling way. We present layout, description, and appropriate similarity computation methods for each one of them. A detailed comparison between these different approaches is given using many experiments conducted on a real dataset. The pros and cons of each approach are illustrated for more advantage that may open a window for future work.

1. Introduction

The huge amount of data and the emerging new ways of marketing enforce the administrators of online systems to search for automatic tools that may facilitate their systems. These systems offer many services to their users ranging from a joke to read or listen to many expensive things to buy online. Literature calls these automatic systems as recommender systems (RSs) with the aim to personalize the user navigation through the Web and direct their action. Today, these systems cover social networks, e-commerce, e-Business, e-Tourist, and many others [1–5]. Recently, Lu et al. [6] reviewed the applications of recommender systems, clustered them into eight main categories and summarized the related recommender system types used for each category.

Formally, there are five phases for building a RS, namely, data collection, user profiling, similarity computation, neighborhood selection, and finally predictions and recommendations. Based on the profile data, RSs can be content-based RSs (CBRSs), collaborative RSs (CRSs), or demographic RSs (DRSs) [1–5]. If the user profile is a set of features extracted from the descriptions of the items user liked before then we have a content-based RS. However, if the user profile is a set of attributes that describe the demographic class or group of the user then we have a DRS. Finally, if the user profile is a list of ratings for items the user has provided before, then we have a CRS which may follow a user-based approach or an item-based approach. Zhang et al. [7] developed a hybrid fuzzy collaborative recommender system which combined user-based and item-based approaches of CRS for mobile products and service recommendations.

The main goal of recommender systems is to address the online information overload problem and to improve the relationship between the system and its customers (users). Both issues are closely related to how the system represents the users and how much processing time is required for fulfilling the customer desire. Among many, DRS is the only system that has a limited number of features that can be fast for thousands if not millions of users. This makes DRS a suitable candidate for many online systems that faces rapid increasing of items and users.

Actually, DRSs do not gain that much popularity due to security and privacy concerns which stand on the top for the user hesitation and the difficulty to obtain true demographic data from the users. However, DRSs are available with a good percentage in our daily life and many online services will be more personalized if this data is taken into account. Age, gender, occupation, income, nationality, and many other demographic data are essential for many applications. For example, age groups are very important when suggesting movies while income ranges are very important when suggesting tourist places. In marketing, male and female shopping requirements are sometimes totally different and we cannot recommend some items without taking the gender of the targeted user into account. Moreover, some RSs suffer from many inherent problems that cannot be solved without hybridization between them and the DRS.

* Correspondence address: Electrical Engineering Department, Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen. Tel.: +967 777796023 (mobile), +966 558137212; fax: +967 4 408068.
E-mail address: mohamad.alshamri@gmail.com

http://dx.doi.org/10.1016/j.knosys.2016.03.006
0950-7051/

This motivates us to explore in some details the profiling approaches of DRS along with the advantages and the appropriate similarity computation methods for each profiling approach. Intuitively, the most important demographic data are age, gender and occupation. Sometimes Zip Code is considered as a demographic data but it is important for some applications only. This makes the research in this field too difficult as the available options are very limited. However, this is not the case if we consider the ways of representing each attribute and the associated similarity methods for comparing them.

This paper studies the user profiling and the similarity computation phases and assumes that the other phases are the same for all approaches. The contributions of this paper are four-fold.

1. Many approaches for profiling users of DRS are studied.
2. We introduce similarity measures for some profiling approaches.
3. We propose a cascaded profiling approach for the neighborhood set generation.
4. We propose a single-attribute profiling approach by treating each attribute as an isolated profile and then merge their predictions.

The rest of this paper is organized as follows: a literature review is given in Section 2. Section 3 is an introduction to DRSs which gives a brief description of DRSs and discusses their advantages and disadvantages. Section 4 introduces the demographic user profile and the nature of the basic attributes for building it. The user profiling approaches for DRSs and the corresponding similarity measures are introduced in Section 5. Section 6 describes the experiments' dataset and the experimental methodology used for examining the profiling approaches. Section 7 discusses the results of the conducted experiments. Finally, we conclude our work in the last section.

2. Literature review

The roots of DRS dated back to 1979 [8], earlier than the notion of recommender system itself in the 90s of the last century [9,10]. Some pure examples for DRS are Grundy [5] which is the first DRS proposal for suggesting books and Lifestyle Finder [11] which aimed to market a range of products and services. To the best of our knowledge, a few number of research papers studied DRS and most of this work was a hybridization between DRS and the other types of RSs for overcoming the weaknesses of these systems like the cold-start problem of the CRSs [12–26].

Al-Shamri and Bharadwaj [12, 13] proposed a compact user model that exploits the user demographic data beside rating-driven features. Demographic data in this model benefited the user and allowed the system to overcome the cold-start problem. Moreover, the added demographic data allows the users with less number of ratings to enjoy the system. A same approach is used in [14] but with a different similarity method. Vozalis and Margaritis [15] proposed a feature combination hybrid RS that used demographic correlations to enhance the prediction accuracy. This work blended different features from different recommendation data sources into a single recommendation algorithm. They used dot product as a similarity measure between the profile vectors. Another work of Vozalis and Margaritis [16] utilized SVD and demographic data at various points of the filtering process in order to improve the predictions quality.

Safoury and Saleh [17] introduced a solution for the cold-start problem by utilizing the demographic data of the new user instead of their ratings. This allowed the system to serve the user even he had no ratings yet. The hybrid RS of Junior et al. [18] employed demographic data to discover and analyze the contextual constraints in a real world recommendations scenario. Ghazanfar and Prugel-Bennett [19] proposed a cascading hybrid RS that combined CBRS, CRS, and DRS. This approach somehow lets each RS to compensate the weaknesses of the others. The importance of demographic data for a research paper RS is studied by Beel et al. [20].

Sobecki [21] proposed two consensus-based hybrid RSs that mixed CBRS, CRS, and DRS at some way. The first proposal mixed CRS and DRS with some contents of the items while the second proposal mixed demographic, collaborative and content-based approaches at different components of the user model. Pazzani [22] proposed an approach that combines recommendations from multiple sources. Traveler agent [23] combined CBRS, CRS, and DRS to bring to the light the positive aspects of each recommender system. Moreno et al. [24] proposed SigTur/E-Destination for tourism and leisure activities using ontologies for guiding the reasoning process.

Said et al. [25] extended CRS to include some demographic features. They argued that these features hold implicit information about users taste and interest. Lu et al. [26] proposed a hybrid fuzzy semantic recommender system that combined item-based fuzzy semantic recommender system and fuzzy item-based fuzzy collaborative recommender system. This hybridization overcomes the semantic limitations of classical collaborative recommender system.

Yujie et al. [27] used the demographic data of new user within a social network to find similar users for him. Chen and He [28] proposed a system that generates user demographic vector from the user information and then employs number of common terms and term frequency for similarity computation. Weber and Castillo [29] studied the effect of some demographic data on the online searching behavior of US people and described how different segments of the population differ in their searching behavior. They argued that revealing the hidden relation between the demographic data and the query type might improve Web search relevance and provide better query suggestions.

3. Demographic recommender systems

DRS is a stereotypical system as it categorizes users based on their demographic attributes. Later, DRS uses the user opinions for the items of the system as a basis for recommendations. It is worth noting that both DRS and CRS utilize user-to-user correlations but based on different data. Therefore the advantages of DRS are almost similar to that of CRS in terms of their unique capacity in identifying cross-genre niches, enticing the users to jump outside the familiar, and their ability to improve themselves over time [3,5].

Formally, DRS has M users, $U = \{u_1, \ldots, u_M\}$, having N demographic attributes, $D = \{a_1, \ldots, a_N\}$. Usually, DRS collects demographic attributes during the registration process using questionnaire about the user demographic data and the user's characteristics [4,5]. Through interacting with the system, the user is asked explicitly or implicitly to rate K items, $S = \{s_1, \ldots, s_K\}$, such as news, Web pages, books, movies, or CDs. Initially, each user u_i may rate a subset of items S_i. The declared rating if available of user u_c for an item s_k is denoted by $r_{c,k}$ [2,10].

After constructing the user profile, DRS calculates the similarity value between the current active user and the remaining training users using a suitable similarity measure. This value indicates how closely the two users in consideration resemble each other. Accordingly, a set of neighbors is selected for this active user from the ranked list of the training users. After that DRS assigns a predicted rating to all the items seen by the neighborhood set and not by the active user. The predicted rating, $pr_{x,\,k}$, indicates the expected interestingness of the item s_k to the user $\mathbf{u}_x$ [2,3]. The predicted rating, $pr_{x,\,k}$, is usually computed as an aggregate of the ratings of