Knowledge-Based Systems 82 (2015) 11–19

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Simultaneous co-clustering and learning to address the cold start problem in recommender systems

Andre Luiz Vizine Pereira [a,b,*], Eduardo Raul Hruschka [a]

[a] University of São Paulo, São Carlos, Brazil
[b] Rubens Lara College of Technology, FATEC-RL, Santos, Brazil

ARTICLE INFO

Article history:
Received 6 September 2014
Received in revised form 17 February 2015
Accepted 18 February 2015
Available online 3 March 2015

Keywords:
Recommender system
Cold starting
Co-clustering
Predictive modeling

ABSTRACT

Recommender Systems (RSs) are powerful and popular tools for e-commerce. To build their recommendations, RSs make use of varied data sources, which capture the characteristics of items, users, and their transactions. Despite recent advances in RS, the cold start problem is still a relevant issue that deserves further attention, and arises due to the lack of prior information about new users and new items. To minimize system degradation, a hybrid approach is presented that combines collaborative filtering recommendations with demographic information. The approach is based on an existing algorithm, SCOAL (Simultaneous Co-Clustering and Learning), and provides a hybrid recommendation approach that can address the (pure) cold start problem, where no collaborative information (ratings) is available for new users. Better predictions are produced from this relaxation of assumptions to replace the lack of information for the new user. Experiments using real-world datasets show the effectiveness of the proposed approach.

1. Introduction

Recommender Systems (RSs) are important components for e-commerce systems [29]. More recently, RSs have also been used to recommend movies [6], songs [17], videos [12], research resources in digital libraries [40,41], and people one may know from social networks [10,32]. To build their recommendations, RSs use varied data sources, which define the characteristics of items, users, and their transactions, and are categorized by the data sources and techniques used, such as, Content Based Filtering (CBF), Demographic Filtering (DF), and Collaborative Filtering (CF). CBF based RSs analyze a set of documents and/or descriptions of items previously evaluated by the user to build a user model or profile, which will then be used by the system for future recommendations of new items [4,38,51]. DF based RSs use attributes such as age, sex, occupation, and educational level to construct a demographic profile of the user, and different recommendations are generated for different demographic niches [30,37]. CF based RSs select items (documents, films, etc.) based on opinions that other users associate with them. Users share information and views about particular items, assigning scores that serve as reference benchmarks to other users.

CBF, DF, and CF RSs have been available for several years, and their advantages, performances, and limitations are well understood [1]. In particular, several CF based systems have been proposed and evaluated in recent years, providing satisfactory results in various commercial applications [9,43,25,34,26]. On the other hand, rather than adopting CF, DF, and CBF as standalone approaches, hybrid systems can combine the strengths better represent user needs [50,4,11,49,39,57,5].

An important issue for RSs is the cold start problem [45,28,2], which occurs due to the lack of prior information about new users and items. For instance, historical data about user profiles and their shopping preferences may not be available [1,42]. Similarly, characteristics of items might be unknown [21,36]. In many situations, the cold start problem leads to loss of new users who decide to stop using the system due to the low accuracy in the first recommendations made by the RS [8]. To minimize the system degradation caused by cold starting, a hybrid approach is proposed that combines collaborative filtering recommendations with demographic information and implements an iterative divide and conquer approach interleaving clustering and learning tasks to construct prediction models. An existing algorithm, SCOAL (Simultaneous Co-Clustering and Learning) [16], is used.

Our contributions are as follows: (i) we propose a hybrid RS based on SCOAL that addresses the (pure) cold start problem, where no collaborative information is available for new users; (ii) we show that better predictions can be built by relaxing the

* Corresponding author at: University of São Paulo, São Carlos, Brazil.
E-mail addresses: avizine@gmail.com (A.L. Vizine Pereira), erh@icmc.usp.br (E.R. Hruschka).

http://dx.doi.org/10.1016/j.knosys.2015.02.016
0950-7051/

assumption on lack of user information. Although this is the expected behavior, quantitative analyses are not common in the literature and, as such, are a complementary contribution to our work.

Section 2 discusses recent advances related to the cold start problem. Section 3 reviews the SCOAL algorithm, and Section 4, presents our approach, based on the SCOAL algorithm, to address the cold start problem. The proposed approach is experimentally evaluated in Section 5, and compared to existing techniques. Section 6 reports our main conclusions and provides some directions for future work.

2. Related work

Many approaches have been proposed to address the cold start problem. A probabilistic model for combining CF and CBF has been proposed by Schein et al. [45]. It uses latent variables, i.e., aspect models between users/movies and movies/actors associations, to make recommendations for new movies. The well-known Expectation Maximization (EM) algorithm [13] is used to estimate the parameters of the model. A hybrid RS proposed by Wang and Wang [53] exploits the probabilistic latent semantic analysis [27] to model the relationship between user attributes and item attributes, and the EM algorithm to fit the model. Gantner et al. [18] use a method that maps item or user attributes to the latent features of a matrix factorization model. With such mappings, one can tackle the new-user and the new-item problems, thereby enhancing speed and predictive accuracy. In an RS based on trust networks [52], users are connected by means of trust scores and receive recommendations for items rated by people belonging to a web of trust network. In a related approach, a recommendation algorithm based on tripartite graphs (users-items-tags) Zhang et al. [58] considers social tags as a bridge between users and items. Tags based on social systems enable users to employ arbitrary tags to label the items of interest. In this setting, predictions are based on both the frequency of tags (such as personal preferences) and the semantic relationships between tags and items (such as global information). Another algorithm based on social networks data was proposed by Sahebi & Cohen [44]. Information from social networks in different dimensions (e.g. interaction between users, ratings, level of popularity/friendship, etc.), is used to detect latent communities of users. Thus, a new user can be inserted into a community that is more similar to his/her profile. The RS can then make recommendations based on ratings from users of that community, which are the closest according to the user's profile. Guo [22,23] merges the ratings of trusted neighbors and thus form a new rating profile for the active users. This new rating profile is then assessed through a Bayesian similarity measure that takes into account both direction and length of rating profiles.

A similarity measure, proximity impact popularity (PIP), was proposed by Ahn [2] as part of his approach to address cold start problems. PIP computes similarities between users and employs these to make recommendations. This approach has shown superior accuracy compared to other measures widely used in CF based RSs, particularly when a small number of ratings is available for computing similarities. Bobadilla et al. [8] proposed a similarity measure that considers not only the numerical information contained in the ratings, such as measures traditionally used in CF, e.g. Pearson correlation, but also uses information about the distribution and number of ratings obtained by each pair of users to be compared. The authors argue that it is more reasonable to set greater similarity between users who have positively evaluated a similar number of items than between users for which the number of items is very different. Basiri et al. [3] apply all the available information for each user to create an ordered weighted averaging operator [56] that is used to make recommendations. The operator uses a set of weights associated with each recommendation technique (CF, CBF, and DF) and their possible combinations to make predictions. Lin et al. [33] used information extracted from Twitter data to make recommendations of mobile applications. Their approach is based on latent Dirichlet allocation [7] which generates latent groups. A new user is mapped to previously defined latent groups using transitive relationships between them and applications.

A hybrid fuzzy linguistic RS based on the quality of items was proposed by Tejeda-Lorente et al. [50]and uses a recommendation strategy that switches between CBF and CF approaches to share user experiences in a university digital library. With this dual perspective, the cold start problem is minimized, because the system switches from one approach to another according to the situation of the system at any given moment.

Shaw et al. [47] use association rules to expand system user profiles based on patterns and associations of items, topics, and categories, thereby giving more information to a recommender system. An algorithm that combines association rules and data clustering has been presented by Sobhanam & Mariappan [48]. The association rules are used to create and expand the user profile to increase the number of ratings made by the user, thus minimizing the cold start problem. Clustering is used to group items that later are used to make predictions for new items. Another approach based on association rules proposed by Leung et al. [31] makes use of cross-level association rules (CLARE) to integrate content information about domain items into collaborative filters. The CLARE algorithm operates on a preference model, comprising user-item and item-item relationships, and infers user preferences for items from the attributes they possess using associations, item attributes, and other domain items, when no recommendations for that item can be generated using CF.

In the predictive model proposed by Park & Chu [35], attributes of movies (e.g. year of production, genre, cast, etc.) and user characteristics (e.g. demographic information, and history behavior) are used as independent variables for linear regression.

In summary, there are different approaches to address the cold start problem. Prediction models typically make use of information about the characteristics of the items and user demographics. There is also a need to identify clusters of users and similar items to improve prediction accuracy. The next section presents the SCOAL algorithm [16], which is a key component of our approach. This algorithm explores evaluations made by users (CF setting), as well as attributes that characterize items and users (CBF and DF settings), to learn a set of models that best represent similar clusters of users and items. Such clusters are automatically identified as part of the model learning process.

3. Review of the SCOAL algorithm

The SCOAL algorithm [16,14] implements an iterative divide and conquer approach that interleaves clustering and learning tasks. A mathematical optimization problem is formulated and an objective function is minimized through an iterative process, until a local minimum is found. A prediction model is generated for each co-cluster using sets of attributes comprising the rows and columns of a matrix, $\mathbf{Y}$, where, for example, rows represent users and columns represent items (see Table 1). In this setting, the similarity between users is indirectly calculated by considering the predictions made by the learning models, and there is one learning model per co-cluster. For regression problems, such as those addressed in this paper, data clustering is based on the continuous outputs of the prediction models. This simultaneous