Expert Systems with Applications 41 (2014) 2065–2073

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Facing the cold start problem in recommender systems

Blerina Lika, Kostas Kolomvatsos *, Stathes Hadjiefthymiades

Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece

ARTICLE INFO

Keywords:
Recommender systems
Cold start problem

ABSTRACT

A recommender system (RS) aims to provide personalized recommendations to users for specific items (e.g., music, books). Popular techniques involve content-based (CB) models and collaborative filtering (CF) approaches. In this paper, we deal with a very important problem in RSs: The cold start problem. This problem is related to recommendations for novel users or new items. In case of new users, the system does not have information about their preferences in order to make recommendations. We propose a model where widely known classification algorithms in combination with similarity techniques and prediction mechanisms provide the necessary means for retrieving recommendations. The proposed approach incorporates classification methods in a pure CF system while the use of demographic data help for the identification of other users with similar behavior. Our experiments show the performance of the proposed system through a large number of experiments. We adopt the widely known dataset provided by the GroupLens research group. We reveal the advantages of the proposed solution by providing satisfactory numerical results in different experimental scenarios.

1. Introduction

Recommender systems (RSs) technology currently is in use in many application domains. RSs can suggest items of interest to users based on their preferences. Such preferences could be retrieved either explicitly or implicitly. In general, recommendations are based on models built from item characteristics or users' social environment. For example, recommendations could be based on preferences of other users having similar characteristics (e.g., age, occupation). The recommendation result is the outcome of a complex process that combines the attributes of items and information about users. Recommendation algorithms try, through intelligent techniques, to identify possible connections between items and users and give the most efficient results. The final aim is the maximization of the quality of recommendation (QoR). As QoR could be defined the value of the matching between a specific item and a specific user.

In literature, one can find the following techniques adopted in RSs: (a) *Collaborative filtering (CF) methods* (Tso-Sutter, Marinho, & Schmidt-Thieme, 2008; Das et al., 2007; Sarvar et al., 2001; Schafer, Frankowski, Herlocker, & Sen, 2007; Jambor & Wang, 2010; Jin, Si, & Zhai, 2006; Wang, de Vries, & Reinders, 2006; Rocchio, 1971; Khabbaz & Lakshmanan, 2011; Popescul, Ungar, Pennock, & Lawrence, 2001) and (b) *Content-based (CB) methods* (Pazzani & Billsus, 2007; Lops, de Gemmis, & Semeraro, 2011; De Gemmis, Lops, & Semeraro, 2007; Middleton, Shadbolt, & De Roure, 2004; Billsus & Pazzani, 2000; Mooney & Roy, 2000). CF systems try to retrieve the final recommendation result through community preferences. Usually, in such systems demographics or user attributes are neglected (Schein, Popescul, Ungar, & Pennock, 2002). More specifically, CF approaches recommend items to a target user based on given ratings by other users in the community. Many algorithms have been proposed for the calculation of similarities between users or items. The selection of the algorithm plays an important role to the final QoR. CB systems try to match user profiles against items description. CB approaches require ratings made by the user herself in contrast to CF models that cannot derive an efficient result without the ratings of other users. Additionally, *Hybrid methods* have been proposed (Popescul et al., 2001; Schein et al., 2002) in order to cover the disadvantages of CF and CB models. These methods combine both techniques in order to provide a more efficient result. Many promising algorithms were presented in the above categories, however, some issues are still open.

One of the most known problems in RSs is the cold start problem. The cold start problem is related to the sparsity of information (i.e., for users and items) available in the recommendation algorithm. The provision of a high QoR in cold start situations is a key challenge in RSs (Park & Chu, 2009). Three types of cold start problems could be identified: (a) recommendations for new users, (b) recommendations for new items, and (c) recommendations on new items for new users. Researchers try to overcome the discussed problem, however, they are interested mainly in item side cold start problems (Zhang, Liu, Zhang, & Zhou, 2010). In this paper, we focus on solving the user side cold start problem. We consider the scenario where a new user

* Corresponding author. Tel.: +30 7275127.
E-mail addresses: mop10319@di.uoa.gr (B. Lika), kostasks@di.uoa.gr (K. Kolomvatsos), shadj@di.uoa.gr (S. Hadjiefthymiades).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.005

asks for recommendations and no data are available for her preferences. Such data are related to ratings for items. Ratings are very important as they show the preferences of a specific user. Additionally, no historical data are present. We propose an algorithm which results the final outcome through three phases. The first phase is responsible to provide means for the classification of the new user in a specific group. For the classification, we adopt efficient techniques like the C4.5 algorithm (Kotsiantis, 2007) and the Naive Bayes algorithm (Zhang, 2004). In the second phase, the algorithm utilizes an intelligent technique for finding the 'neighbours' of the new user. We examine important characteristics of the user and try to find other users inside the group that best match to her. In the third phase, the final outcome is calculated. This is done adopting prediction techniques for estimating the ratings of the new user. In comparison with research efforts found in the literature, our work has the following differences. Our model:

- handles the new user cold start problem,
- does not require any a priori probability to be known like efforts adopting probabilistic models,
- does not require any interview process,
- does not depend on any complex calculations,
- involves semantic similarity metrics in the calculation process.

The structure of the paper is as follows. Section 2 describes important research efforts in the domain of RSs and the cold start problem. Section 3 gives a high level description of the proposed system while Section 4 presents in detail the key components of our RS. Section 5 is devoted to the presentation of evaluation metrics and the description of our experimental results. Finally, Section 6 concludes the paper.

2. Related work

The CF methods are categorized to (Khabbaz & Lakshmanan, 2011): (i) user-based, (ii) item-based, (iii) model-based, and, (iv) fusion-based approaches. In the user-based approaches (Herlocker, Konstan, Borchers, & Riedl, 1999), a similarity matrix is adopted to store the ratings of each user for every item. The main problem is when missing values are present. The item-based methods (Deshpande, 2004; Sarvar et al., 2001; Wang et al., 2006) adopt pairwise item similarities which are more reliable than user similarities, thus, resulting in higher QoR. The model-based methods (Das et al., 2007; Canny, 2002; Jin et al., 2006) exploit the sparsity of data in the similarity matrix. Training examples are used to generate the appropriate model parameters. Based on such parameters, missing values could be substituted. However, tuning a significant number of parameters has prevented these methods from wide adoption (Jambor & Wang, 2010). The fusion-based methods (Tso-Sutter et al., 2008; Zhang et al., 2010; Oku & Hattori, 2011) adopt information fusion techniques for building the final items list.

As mentioned, CB systems try to match user profiles against items description. Various techniques have been used in CB models like keyword-based models (Asnicar & Tasso, 1997; Chen & Sycara, 1998; Mladenic, 1999; Moukas, 1997), semantic techniques (Basile, de Gemmis, Gentile, Iaquinta, & Lops, 2007; De Gemmis et al., 2007; Eirinaki, Vazirgiannis, & Varlamis, 2003; Magnini & Strapparava, 2001; Middleton et al., 2004) or probabilistic models (Billsus & Pazzani, 1999; Billsus & Pazzani, 2000; Mooney & Roy, 2000; Pazzani & Billsus, 1997). Keyword-based models handle every document as a vector where each dimension describes a specific term. Weights are used to define the association between documents and terms (i.e., user profiles and item characteristics). Semantic techniques provide means for reasoning in the recommendation mechanisms. Ontologies play an important role to that, however, one can identify the problem of heterogeneity. Different item providers could utilize different ontologies, thus, the reasoning process becomes very hard. Probabilistic methods yield posterior probabilities by analysing historical data and based on a priori probabilities. Such probabilities are related to the relationship between documents and terms. Usually, the Bayes rule is the key methodology for the calculation of such probabilities while the Naive Bayes classifier is recognized as the method with the best performance (Lewis & Ringuette, 1994).

The adoption of CB systems has a number of advantages and disadvantages (Lops et al., 2011). These approaches require ratings made by the user herself in contrast to CF models that cannot derive an result without other users ratings. However, in CB systems the cold start problem is intense as ratings are not available for new users. CB models depend on the performance of content analysis methodology they adopt. Explanations on the final result could be given by terms of items descriptions and users profiles something that cannot be done in CF approaches. Additionally, new items are handled easier as the recommendation is based on their descriptions even if ratings are not present yet. The performance of the matching process of the item descriptions with the user profiles is also a critical issue.

A number of research efforts deal with the cold start problem. The combination of collaborative data and content is proposed as a solution to the discussed problem (Popescul et al., 2001; Schein et al., 2002). Such models incorporate three data sources: users, items and item contents. The influence of collaboration data with content emerges naturally from the given data sources by adopting a probabilistic model. Six techniques that CF systems can use to learn about new users are presented in Al Mamunur et al. (2002). These techniques select a sequence of items to present to every new user. In Massa and Bhattacharjee (2004), the authors try to assert the cold start problem through a trust-aware system that takes into account the 'web of trust' provided by every user. The proposed model involves trust propagation between users and inference on the weights of unknown users. A recommendation algorithm based on social tags is proposed in Zhang et al. (2010). The algorithm provides personalized recommendations especially when the assigned tags belong to diverse topics. In Lam, Vu, Le, and Duong (2008), an hybrid approach is discussed. The proposed model utilizes a combination of the CF approach with the CB. Two probabilistic aspect models using pure CF try to handle the new user problem. The use of association rules is the proposed solution in Shaw, Xu, and Geva (2010). Through such rules, the authors try to expand the user profile and, thus, avoid the cold start problem. The performance is improved using non-redundant rule sets. However, complete rule enumeration is often intractable for datasets with a very large number of multi-valued attributes. In Zhou, Yang, and Zha (2011), the authors present functional matrix factorization (fMF). fMF constructs a decision tree for the initial interview (each node being an interview question) enabling the RS to query the user adaptively. Hence, the interview phase could be 'alleviated' thus improving the performance of the model. In Park and Chu (2009), the authors propose predictive feature-based regression models that leverage all the available information of users and items to tackle the cold-start problem. Finally, in Golbandi, Koren, and Lempel (2011), a model for the profiling of new users is discussed. The proposed model is a kind of an interview that elicits the opinion of users about items. The model involves an adaptation scheme on the users' answers in order to provide a more efficient result.

3. The proposed model

The proposed model alleviates the user cold start problem of CF. The main operational aspects are depicted in Fig. 1. The process of