Neurocomputing 176 (2016) 72–80

Contents lists available at ScienceDirect

Neurocomputing

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

Web mining based framework for solving usual problems in recommender systems. A case study for movies' recommendation

María N. Moreno *, Saddys Segrera, Vivian F. López, María Dolores Muñoz, Ángel Luis Sánchez

Department of Computing and Automation. University of Salamanca, Plaza de los Caídos s/n, 37008 Salamanca, Spain

ARTICLE INFO

Article history:
Received 9 March 2014
Received in revised form
8 October 2014
Accepted 29 October 2014
Available online 9 May 2015

Keywords:
Recommender systems
Semantic web mining
Associative classification
First-rater
Cold-start
Sparsity

ABSTRACT

Endowing web systems with efficient and reliable recommendation procedures is being the target of intensive research in the last years. Numerous methods have been proposed to provide users with more and more effective recommendations, from the traditional collaborative filtering approaches to sophisticated web mining techniques, however some important drawbacks are still present in current recommender systems. Some works in the literature address these problems in an individual way. In this work, we propose a complete framework to deal jointly with some of the most important: scalability, sparsity, first rater and cold start problems. Although the framework is addressed to movies' recommendation and validated in this context it can be easily extended to other domains. It manages different predictive models for making recommendations depending on specific situations. These models are induced by data mining algorithms using as input data both product and user attributes structured according to a particular domain ontology.

1. Introduction

Web recommender systems are used in many application domains to predict consumer preferences and assist web users in the search for products or services. Recommender mechanisms implemented in current systems have different levels of complexity, ranging from those that recommend products based on associations between them in previous transactions, to those that make recommendations based on evaluations that users provide about products and similarity between user preferences. The latter, known as collaborative filtering (CF) methods, are the most successful; however, some important drawbacks in them have been reported [35,31], especially in traditional approaches based on nearest neighbor algorithms, which show serious performance and scalability problems. In addition, the great number of evaluations needed by these methods in order to provide precise recommendations causes the sparsity problem when evaluations from users are insufficient. Improvements deriving from the research carried out over several years are being incorporated to current web systems, yielding more and more effective recommendations. Data mining algorithms have been applied to deal with sparsity and performance problems since they are not only based on product evaluations but on other attributes. Moreover, they are induced offline, before the user logs onto the system, and therefore the time spent on building the model has no effect on the user response time. Nevertheless, sparsity can also reduce the precision of data mining by different degrees depending on the type of algorithm. Therefore, it is necessary to find data mining algorithms slightly sensitive to sparsity in order to obtain precise recommendations.

On the other hand, in spite of the advantages of data mining methods, there are situations in recommender systems in which it is very difficult to give recommendations to the user. For instance, when new products without evaluations are introduced into the catalog or when a new user without evaluations about products requests recommendations, the first-rater and the cold-start problems arise respectively [14,18].

In this paper, a recommendation framework is proposed, which aims at overcoming the main drawbacks of current recommender systems. It constitutes a semantic based web mining approach that is applied in the context of a movies' recommender system but it can be easily extended to other application domains. Recommendations are obtained by means of applying associative classification to data annotated with semantic metadata according to a domain-specific ontology.

The rest of the paper is organized as follows: Section 2 is devoted to describing the state of the art and the main problems of recommender systems. In Section 3 the proposed framework is presented. The framework is validated through a case study reported in Section 4, where the ontology for this particular application and a comparative study of the results from different algorithms are also gathered. Finally, the conclusions are given in Section 5.

* Corresponding author. Tel.: +34 923 294400x1513.
E-mail address: mmg@usal.es (M.N. Moreno).

http://dx.doi.org/10.1016/j.neucom.2014.10.097
0925-2312/

2. Related work

2.1. Recommender systems' methods

Recommendation methods can be classified into two main categories [23]: collaborative filtering (CF) and content-based approach. Techniques in the first category, initially based on nearest neighbor algorithms, are used to predict product preferences for a user based on the opinions of other users. The opinions can be obtained explicitly from the users as a rating score or by using some implicit measures from purchase records as timing logs [34]. The content based approach was first used in the information recovery field in order to recommend text documents by comparing between their contents and user profiles. The weights for the words extracted from the document are added to the weights for the corresponding words in the user profile, if the user is interested in the page [23]. The main shortcoming of this approach is the lack of mechanisms to manage web objects such as motion pictures, images, music, etc. Besides, it is very difficult to handle the big number of words obtained from the product contents. This approach has been extended to any kind of products by replacing document words by product attributes.

Currently there are two approaches for collaborative filtering, memory-based (user-based) and model-based (item-based) algorithms. Memory-based algorithms, also known as nearest-neighbor methods, were the earliest used [32]. They treat all user items with statistical techniques in order to find users with similar preferences (neighbors). The prediction of preferences (recommendations) for the active user is based on the neighborhood features. Several similarity functions can be applied although the most popular is Pearson correlation coefficient. Once the most similar items are found the prediction is computed by taking a weighted average of the product ratings of the nearest neighbors. The advantage of these algorithms is the quick incorporation of the most recent information, but the disadvantage is that the search for neighbors in large databases is slow [35]. In order to avoid this inconvenience, model-based CF algorithms have been proposed. They use data mining techniques in order to develop a model of user ratings, which is then employed to predict user preferences. There are a great variety of data mining algorithms that can be applied in model-based CF. Neural networks were the first of this kind of method [7], which changed the nearest neighbor approach of CF methods for a classification approach. Bayesian networks constitute another technique widely used in the induction of recommendation models in a single way [11], or jointly with other methods [12]. The main shortcoming of these methods is the high computational cost of building the net, especially when the amount of data is high. Support Vector Machines (SVM) can also be used in recommender systems [42]. In some works, SVM is used as a complementary technique for other methods [15].

The works referenced in this section are just a small sample of the numerous data mining proposals to be used in collaborative filtering based recommender systems. However, the current trend, especially in sparse contexts where ratings are insufficient, is to exploit hybrid methodologies combining content-based and collaborative filtering approaches in order to take advantage of the strengths of each of them [6]. In recent works, semantic information is added to the available data in order to formalize and classify product and user features. These works are commented in Section 2.3.

Collaborative filtering (CF), especially the memory-based approach, has certain limitations that have an important impact on the quality of the recommendations. First, rating schemes can only be applied to homogeneous domain information. Furthermore, sparsity and scalability are serious weaknesses which would lead to poor recommendations [13]. Sparsity occurs when the number of ratings needed for prediction is greater than the number of the ratings obtained because CF usually requires user-explicit expression of personal preferences for products. The second limitation is related to performance problems in the search for neighbors in memory-based algorithms. These problems are caused by the need to process large amounts of information. The computer time grows linearly with both the number of customers and the number of products in the site. The lesser time required for making recommendations is an important advantage of model-based methods. This is due to the fact that the model is built off-line before the active user goes into the system, but it is applied on-line to recommend products to the active user. Therefore, the time spent in building the model has no effects on the user response time since little processing is required when recommendations are requested by the users, contrary to the memory-based methods that compute correlation coefficients when the user is on-line. Nevertheless, model-based methods present the drawback that recent information is not added immediately to the model but a new induction is needed in order to update the model.

The quality of the recommendations for the users has an important effect on the clients' retention. Users refuse poor recommender systems which can cause two types of error: false negatives, which are products that are not recommended, though the customer would like them, and false positives, which are products that are recommended, though the customer does not like them [13]. The most serious errors are false positives, because these errors will cause negative reactions in the customers and thus they will not probably visit the site again. The use of data mining algorithms to find customers characteristics that increase the probability of buying recommended products can help to avoid these problems.

Although the drawbacks described above may be minimized by means of data mining algorithms, there are other shortcomings that may occur even with these methods. The first-rater (or early-rater) problem arises when it is not possible to offer recommendations about an item that was just incorporated into the system and, therefore, has few evaluations (or even none) from users. In fact, the early rater problem is directly linked to sparsity since when a system has a high number of items, probably most of these items have never received any evaluation. Conceptually, the early-rater problem can be viewed as a special instance of the sparsity problem [20]. Sarwar et al. [34] affirm that current recommender systems depend on the altruism of a set of users who are willing to rate many items without receiving many recommendations. Economists have speculated that even if rating required no effort at all, many users would choose to delay considering items to wait for their neighbors to provide them with recommendations [4]. Thus, it is necessary to find a way to encourage users to make evaluations about items available in the system.

Analogously, this drawback also occurs with a new user joining the system: since there is no information about his preferences, it would be impossible to determine his behavior in order to provide recommendations. Actually, this variant of the first-rater problem is also referred to as the “cold-start problem” [18] in the literature. The grey-sheep problem [14] is another drawback associated with collaborative filtering methods. This problem refers to the users who have opinions that do not consistently agree or disagree with any group of users. As a consequence, such users do not receive recommendations.

The problems addressed here have been treated in some works in the literature. One way of dealing with sparsity and scalability problems consists of reducing the dimensionality of the database used for CF by means of a technique called Singular Value Decomposition (SVD) [40]. Barragáns-Martínez et al. [6] have adapted the proposal of Vozalis and Margaritis for a hybrid system combining content-based and CF approaches in the TV program recommendation domain.

The association of users and items with a set of concepts is other approach used to improve the precision of recommendation and to solve the sparsity problem. In [24] a probabilistic learning framework for collaborative learning recommendation is presented. Law of total