International Journal of Approximate Reasoning 51 (2010) 785–799

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks ☆

Luis M. de Campos, Juan M. Fernández-Luna *, Juan F. Huete, Miguel A. Rueda-Morales

Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I. Informática y de Telecomunicación, CITIC-UGR Universidad de Granada, C.P, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 15 February 2008
Received in revised form 5 April 2010
Accepted 5 April 2010
Available online 11 April 2010

Keywords:
Content-based filtering
Collaborative filtering
Hybrid recommender systems
Bayesian networks
MovieLens
IMDB

ABSTRACT

Recommender systems enable users to access products or articles that they would otherwise not be aware of due to the wealth of information to be found on the Internet. The two traditional recommendation techniques are content-based and collaborative filtering. While both methods have their advantages, they also have certain disadvantages, some of which can be solved by combining both techniques to improve the quality of the recommendation. The resulting system is known as a hybrid recommender system.

In the context of artificial intelligence, Bayesian networks have been widely and successfully applied to problems with a high level of uncertainty. The field of recommendation represents a very interesting testing ground to put these probabilistic tools into practice.

This paper therefore presents a new Bayesian network model to deal with the problem of hybrid recommendation by combining content-based and collaborative features. It has been tailored to the problem in hand and is equipped with a flexible topology and efficient mechanisms to estimate the required probability distributions so that probabilistic inference may be performed. The effectiveness of the model is demonstrated using the MovieLens and IMDB data sets.

1. Introduction

Recommender systems (RSs) attempt to discover user preferences, and to learn about them in order to anticipate their needs. Broadly speaking, a recommender system provides specific suggestions about items (products or actions) within a given domain, which may be considered of interest to the given active user [1]. Formally, in a hybrid recommending framework, there exists a large number m of items or products $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$, which are described by a set of l attributes or features, $\mathcal{F} = \{F_1, F_2, \ldots, F_l\}$, and each product is specified by one or several. There is also a large set of n users, $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ and for each user, a set of ratings about the quality of certain observed items in $\mathcal{I}$. Under this formulation we distinguish two different problems:

- Given an item not rated, predicting the rating that the user would give.
- Given a user, find the best items and their ratings for being recommended, showing the results ordered by predicted rating.

☆ Work jointly supported by the Spanish Ministerio de Educación y Ciencia (TIN2008-06566-C04-01) and the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

* Corresponding author.
E-mail addresses: lci@decsai.ugr.es (L.M. de Campos), jmfluna@decsai.ugr.es (J.M. Fernández-Luna), jhg@decsai.ugr.es (J.F. Huete), mrueda@decsai.ugr.es (M.A. Rueda-Morales).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.04.001

Although both notions are closely related, this paper deals with the first type, i.e. rating prediction. The usual formulation of the problem is then to predict how an active user might rate an unseen item.

Many different approaches to the recommender system problem have been published [2–4], using methods from machine learning, approximation theory, and various heuristics. Independently of the technique used and based on how the recommendations are made, recommender systems are usually classified [3] into the following categories: *Collaborative filtering systems* that attempt to identify groups of people with similar tastes to those of the user and recommend items that they have liked and *Content-based recommender systems* which use content information in order to recommend items similar to those previously preferred by the user.

Generally, collaborative systems report a better performance than content-based approaches, but their success relies on the presence of a sufficient number of user ratings [3,5,6,4,7]. Such systems have the drawback that they suffer from the item cold-start problems which occur when recommendations must be made on the basis of few recorded ratings [8,3]. These problems arise because the similarity analysis is not accurate enough. In these situations the use of a content-based approach appears as an alternative. Nevertheless, this approach has its own limitations. For example, the keywords used to represent the content of the items might not be very representative. Also, content-based approaches suffer the limitation of making accurate recommendations to users with very few ratings.

A common approach to solve the problems of the above techniques is to combine both content-based and collaborative information into a *hybrid recommender system* [9]. Different hybridization methods [3,6,9,10] have been proposed, such as the use of weighted criterion (the scores of different recommendation components are combined numerically), the use of a switching mechanism (the system chooses among recommendation components and applies the selected one) or even the presentation of the two recommendations together, leaving the decision in the user's hands. Nevertheless, a common problem with these methods is that the parameters controlling the hybridization have to be tuned.

This is the setting for the proposal presented in this paper, i.e. the design of a hybrid system with the aim of predicting how an active user should rate a given item. Particularly, we will explore the use of Bayesian network formalism to represent the relationships among users $\mathcal{U}$, items $\mathcal{I}$ and features $\mathcal{F}$, the elements involved in the recommendation processes. By using Bayesian networks, we can combine a qualitative representation of how users and items are related (explicitly representing dependence and independence relationships in a graphical structure) as well as a quantitative representation by means of a set of probability distributions, measuring the strength of these relationships.

In our proposal we shall distinguish two different parts: The first one is used to represent the knowledge that we have about how the active user rates the items, i.e. the user profile, which includes both content-based and collaborative information. The second component represents those relationships related to the target item. We would like to say that content-based information is not only used to improve the active user knowledge, but also this information has been used to improve the knowledge at the collaborative level. This is possible because we have a hybrid model where all the components are represented under the same formalism. By means of this fact we can explore the importance of the different elements in the quality of the predictions.

In order to present the model, this paper is structured in the following way. The following section introduces recommender systems and reviews the related work. Section 3 describes the model from a topological point of view. How to use the recommender model and how inference is performed efficiently are explained in Section 4. Section 5 discusses the probability distribution estimation. In order to determine the performance of the proposed model, it is evaluated in Section 6. Finally, Section 7 presents our conclusions and outlines future lines of research.

2. Related work and preliminaries

Based on how the recommendations are made, recommender systems are classified into:

- *Content-based recommender systems* that [3] store content information about each item to be recommended. This information will be used to recommend items similar to those previously preferred by the user, based on how similar certain items are to each other or the similarity with respect to user preferences (also represented by means of a subset of content features). Focusing on probabilistic approaches, learning as a constraint satisfaction problem is considered in [11], where the user profile is learnt by considering contextual independence. By assuming independence between variables, Bayesian classifiers have also been used in [12,13] to estimate the probability of an item belonging to a certain class (relevant or irrelevant) given the item description. Also, Bayesian networks [14,15] have been used to model the item's description.
- *Collaborative filtering systems* [3] attempt to identify groups of people with similar tastes to those of the user and recommend items that they have liked. According to [16], collaborative recommender systems can be grouped into *memory-based* and *model-based* approaches.

On the one hand, memory-based algorithms use the entire rating matrix to make recommendations. In order to do so, they use an aggregation measure by considering the ratings of the other users [17] (those most similar) for the same item. Different models can be obtained by considering different similarity measures and different aggregation criteria. Also *item-based* approaches, which take into account the similarity between items (two items are similar if they have been rated similarly) [18,19], appear as good alternatives to the user-based method.