Electronic Commerce Research and Applications 14 (2015) 542–562

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS

Mehrbakhsh Nilashi [a,*], Othman bin Ibrahim [a,*], Norafida Ithnin [a], Nor Haniza Sarmin [b]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
[b] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form 10 August 2015
Accepted 10 August 2015
Available online 20 August 2015

Keywords:
Expectation Maximization
Clustering
PCA
ANFIS
Multi-criteria CF
Tourist recommendation

ABSTRACT

In order to improve the tourist experience, recommender systems are used to offer personalized information for online users. The hotel industry is a leading stakeholder in the tourism sector, which needs to provide online facilities to their customers. Collaborative Filtering (CF) techniques, which attempt to predict what information will meet a user's needs based on data coming from similar users, are becoming increasingly popular as ways to combat information overload. They use a single rating as input. However, the multi-criteria based CF presents a possibility to provide accurate recommendations by considering the user preferences in multiple aspects and they can be an appropriate choice for the tourist. In this paper, we propose a new hybrid method for hotel recommendation using dimensionality reduction and prediction techniques. Accordingly, we have developed the multi-criteria CF recommender systems for hotel recommendation to enhance the predictive accuracy by using Gaussian mixture model with Expectation Maximization (EM) algorithm and Adaptive Neuro-Fuzzy Inference System (ANFIS). We have also used the Principal Component Analysis (PCA) for dimensionality reduction and to address multi-collinearity induced from the interdependencies among criteria in multi-criteria CF dataset. Our experiments confirmed that the proposed hybrid method achieved high accuracy for hotel recommendation for the tourism sector.

1. Introduction

Tourism as a strategic sector has provided a significant contribution to the economies of many nations around the world (Wan Lee and Brahmasrene 2013). It has provided a remarkable impact on the global economic development in which the contribution of this sector for the employing people and economic activity have estimated around 7.6% of global employment and US$ 5474 billion of 9.4% of global GDP (World Travel and Tourism Council 2009). According to the World Tourism Organization (2006), it is predicted that by 2020, tourist arrivals around the world will increase by over 200%. Impressive changes in the Information and Communications Technologies (ICTs) and the Internet has resulted in extensive transformation of the industry. According to the Travel Industry Association of America (www.tia.org) cited by (Lucas et al. 2013), in 2003, the major United States adult population (around 30%) has used Internet as a tool to check prices and schedules and seek information regarding destinations. 66% of tourists booked travel needs using the Internet. In addition, the ICTs have considerably improved the innovations in the tourism sector in management and marketing of tourism packages and brought about new paradigm shifts in this sector as discussed in many researches (Polo Peña et al. 2013, Chiu et al. 2009, Popescu and Grefenstette 2011, Morrison et al. 2001, Singh and Kasavana 2005, Connolly and Lee 2006, Pan et al. 2007, 2011, Buhalis and Law 2008, Xiang and Pan 2010, Buhalis and O'Connor 2005).

Tourism is an activity closely linked with personal interests and preferences (Chou et al. 2008, Wang et al. 2002, Benítez et al. 2007). Recommender systems designed in the tourism domain and applications known as Travel Recommender Systems (TRSs) or destination recommendation system, are a valuable tool for customers and travel agencies (Loh et al. 2004, Werthner and Ricci 2004). That is why many tourism web applications incorporate recommender systems. With this, they try to simulate the interaction with a human travel agent. Through the introduction of tourism recommending systems, tourists can easily access information about the hotels they need, thus, resulting in shorter lead-time for bookings, making last-minute decisions and generally, tailoring their preferences. Tourism recommender systems are a class of intelligent systems that render tourism related information services in the form of

* Corresponding authors.
E-mail addresses: nilashidotnet@hotmail.com (M. Nilashi), othmanibrahim@utm.my (O. bin Ibrahim).

http://dx.doi.org/10.1016/j.elerap.2015.08.004
1567-4223/© 2015 Elsevier B.V. All rights reserved.

guides and suggestions to users. This class of systems can be broadly classified as web-based tourism recommender systems and mobile recommender systems. Web-based tourism recommender systems are intelligent systems that are usually embedded in e-Tourism portals in order to deliver travel information guide, travel advice and travel planning recommendations.

Collaborative Filtering (CF) can be an appropriate choice for tourism object recommendation. Recommender systems based on CF are those in which recommendations only consider the similarity of terms between users. That is, collaborative systems recommend items that other users with similar interests like. However, traditional CF use a single rating as input, usually an overall numerical ranking by a user to an item. Hence, in some applications, this kind of recommendation does not meet users' personalized needs and multi-criteria ratings are considered.

Multi-criteria based CF presents a possibility of providing accurate recommendations by considering the user preferences in multi aspects of items. According to Adomavicius and Kwon (2007), pure CF-based recommender systems rely solely on product ratings provided by a large user community to generate personalized recommendation lists for each individual online user. In traditional CF systems the assumption is that customers provide an overall rating for the items which they have purchased, for example, using a 5-star rating system. However, given the value of customer feedback to the business, customers in some domains are nowadays given the opportunity to provide more fine-grained feedback and to rate products and services along various dimensions (Jannach et al. 2012a, Adomavicius et al. 2011, Nilashi et al. 2014c). According to Adomavicius and Kwon (2007), multi-criteria system provides more information about user preferences than a single-rating system. And by adopting a decision theory, multi-criteria systems can provide rich tools for system designers to build more interesting systems as well (Lakiotaki et al. 2011). In addition, nowadays, allowing online visitors to provide fine-grained multi-criteria rating feedback is common in the travel and tourism sector. TripAdvisor tourism is one of the popular platforms which has provided users to rate hotels according to different criteria such as cleanliness, service or value for money (Nilashi et al. 2015a).

Adomavicius and Kwon (2007) developed a number of basic strategies to exploit multi-criteria ratings for improving the predictive accuracy of a recommender in terms of typical information retrieval measures. Later on, a number of additional techniques to leverage the detailed ratings in the recommendation process were proposed (Liu et al. 2011, Sahoo et al. 2012, Shambour and Lu 2011a,b, Jannach et al. 2012a,b, 2014, Nilashi et al. 2014a,b, 2015a). The work presented in this paper continues on these lines of research.

Overall, our work is similar to the works of which has used, clustering, combined with methods such as methods developed by Adomavicius and Kwon (2007) and Jannach et al. (2012a,b), where we use Neuro-Fuzzy techniques to predict the overall ratings from the given multi-criteria ratings. Furthermore, our work is, in some sense similar to that of Liu et al. (2011) and Nilashi et al. (2015a) where we apply clustering.

From the literature on multi-criteria CF, at the moment there is no implementation of PCA, Neuro-Fuzzy and clustering recommenders in multi-criteria CF, and this research tries to develop a recommender system in the tourism sector based on these approaches. Thus, in order to improve predictive accuracy of multi-criteria CF, we propose a new model using fuzzy logic, neural networks and clustering techniques. To the best for our knowledge, an artificial intelligence method (ANFIS), clustering method (EM) and dimensionality reduction (PCA) is applied for the first time in this research in the context of multi-criteria CF recommendations in particular for hotel recommendation based on multi-criteria CF.

1.1. Recommendation problem

In multi-criteria CF problem, there are m users, n items and k criteria in addition to the overall rating. Users have provided a number of explicit ratings for items; a general rating R_0 must be predicted in addition to k additional criteria ratings ($R_1,\ldots, R_k$). It can be configured to push new items to users in two ways, either by producing a Top-N list of recommendations for a given target, or by predicting the target user's likely utility (or rating) for a particular unseen item. We will refer to these as the recommendation task and the rating prediction task in multi-criteria CF, respectively. Fig. 1 demonstrates the multi-criteria CF problem in case of prediction and recommendation tasks for an active user U_a and active item I_j.

Recommendation is a list of N products, $TP = \{T_{p1}, T_{p2,}\ldots, T_{pN}\}$, that the active user will like the most. The recommended list usually consists of the products not already purchased by the active customer. This output interface of multi-criteria CF algorithms is also known as Top-N recommendation. Multi-criteria CF algorithms represent the entire $m \times n \times k$ user-item-criteria data as a tensor of ratings, $\underline{A}$. Each entry $a_{i,j}$ in tensor $\underline{A}$ as shown in Fig. 1 represents the preference score (ratings) of the ith user on the jth item (hotel) as overall preference in addition to criteria ratings in the 3rd dimension. Each overall and criteria rating is within a numerical scale and it can as well be 0, indicating that the user has not yet rated that item.

Thus, the algorithm for a multi-criteria recommender system can be extended from a single-rating recommender system. Following this approach, Adomavicius and Kwon presented two approaches to leverage multi-criteria ratings through extending single-rating CF (Adomavicius and Kwon 2007). One is computing the overall user similarity through aggregating the similarities calculated from each individual criterion (Adomavicius and Kwon 2007).The other approach is aiming for a more holistic calculation of user similarity through multidimensional distance metrics. Each rating is presented in a multivariable format, such as

$$r_{u,i} = f(r_0, r_1, \ldots, r_k) \tag{1}$$

where r_0 is the overall rating that user u has rated item i and $r_1,\ldots, r_k$ presents the rating of criterion $1,\ldots, k$.

In this paper, we use Pearson correlation coefficient approach for users and items similarity calculation. The Pearson correlation coefficient (the most commonly used weighting approach) measures the degree to which a linear relationship exists between two variables (McLaughlin and Herlocker 2004). In this research, it is used to evaluate how a certain user is related to an active user with respect to their preferences on given items. The Pearson correlation coefficient is derived from a linear regression model (see Eq. (2)). The range of the result of the equation is from −1 to 1, inclusively. More specifically, a result of "1" means the two users are positively related (absolute agreement), "−1" denotes they are negatively related (absolute disagreement), and "0" indicates no relation at all.

$$sim_{u,v} = \begin{cases} 0 & \begin{array}{l} \textit{if } \sum_{i \in I_{uv}} (r_{u,i} - \overline{r_u})^2 = 0 \\ \textit{or if } \sum_{i \in I_{uv}} (r_{v,i} - \overline{r_v})^2 = 0 \\ \textit{or if } |I| = 0 \end{array} \\ \dfrac{\sum_{i \in I_{uv}} (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \overline{r_v})^2}} & \textit{Otherwise} \end{cases} \tag{2}$$

where $I_{u,v}$ in Eq. (2) is a set of items that both user u and v rate, I_u is a set of items that user u rates and I_v is a set of items that user v rates.