Expert Systems with Applications 41 (2014) 5680–5688

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Power coefficient as a similarity measure for memory-based collaborative recommender systems

Mohammad Yahya H. Al-Shamri *

Department of Electrical Engineering, Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen
Computer Networks and Communications Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia

ARTICLE INFO

Keywords:
Web personalization
Collaborative recommender system
Similarity measure
Jaccard coefficient
Dice coefficient
Power coefficient

ABSTRACT

E-commerce systems employ recommender systems to enhance the customer loyalty and hence increasing the cross-selling of products. However, choosing appropriate similarity measure is a key to the recommender system success. Based on this measure, a set of neighbors for the current active user is formed which in turn will be used later to recommend unseen items to this active user. Pearson correlation coefficient, the most popular similarity measure for memory-based collaborative recommender system (CRS), measures how much two users are correlated. However, statistic's literature introduced many other coefficients for matching two sets (vectors) that may perform better than Pearson correlation coefficient. This paper explores Jaccard and Dice coefficients for matching users of CRS. A more general coefficient called a Power coefficient is proposed in this paper which represents a family of coefficients. Specifically, Power coefficient gives many degrees for emphasizing on the positive matches between users. However, CRS users have positive and negative matches and therefore these coefficients have to be modified to take negative matches into consideration. Consequently, they become more suitable for CRS research. Many experiments are carried out for all the proposed variants and are compared with the traditional approaches. The experimental results show that the proposed variants outperform Pearson correlation coefficient and cosine similarity measure as they are the most common approaches for memory-based CRS.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The fast growing Web drives e-commerce at a very fast pace. Today, people are moving from surfing the Web to shopping online at a faster pace than ever before. E-commerce allows customers to shop during off-hours, save time, and avoid going to the stores. However, to entice customers, e-commerce systems have to personalize customers' navigation and therefore introducing them many unseen products. This is done by so called recommender systems which become a must in many Web applications. Such systems, for example, offer customers of on-line retailers suggestions about what they might like to buy, based on their past history of purchases. Amazon, eBay, MovieLens and many others use some type of recommender systems (Adomavicius & Tuzhilin, 2005; Bobadilla, Ortega, Hernando, & Gutiérrez, 2013; Miller, Albert, Lam, Konstan, & Riedl, 2003).

The most successful recommender system is the collaborative recommender system (CRS) (Miller et al., 2003) which recommends items to a given user based on the opinions of his like-mined neighbors who have more experience on a given topic. Day-by-day, CRS becomes a must Web personalization tool to confront the information overload problem which affects our everyday experience while searching for information from the Web. Today many Web applications use CRS to personalizes the customers experience with their neighbors. These applications range from movies, music, learning, products, and tourist locations. The main advantage of CRS is that it recommends out-of-the-box items to an active user based on his neighbors past tastes and preferences (Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2013; Rajaraman, Leskovec, & Ullman, 2012; Schafer, Frankowski, Herlocker, & Sen, 2007).

Normally, we cannot expect users come across or have heard of each of the products they might like. Thus the set of neighbors play an important role to any CRS success. As long as this set is close and representative to the active user, the CRS suggestions will be more valuable. Hence, a great attention has to be paid for selecting this set of neighbors.

* Address: Department of Electrical Engineering, Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen. Tel.: +967 777796023, +966 558137212; fax: +967 (4) 408068.
E-mail address: mohamad.alshamri@gmail.com

http://dx.doi.org/10.1016/j.eswa.2014.03.025
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Formally, CRS constructs a set of neighbors for an active user based on a similarity measure which has to reflect the most important factors between the two users in consideration. Consequently, the similarity computation phase for any CRS plays an important role for its success. Different similarity measures often lead to different sets of neighbors for a given active user. The current most common similarity measures for memory-based CRS are Pearson correlation coefficient and cosine similarity measure (Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2013; Rajaraman et al., 2012; Schafer et al., 2007). However, many statistical coefficients can be used as similarity measures for CRS. The ultimate goal is to get a set of neighbors that are close as possible to the given active user so that the system accuracy is enhanced. The cosine similarity measure relies on the users' raw declared ratings while Pearson correlation coefficient relies on the deviation of these ratings from the user's mean of ratings. Both of them aggregate the whole ratings or deviation to reach the final result.

Similarity measures are crucial for selecting the best neighbors for a given active user. So far, most the CRS similarity measures are symmetric treating both positive (liked items) and negative matches (disliked items) equally. However, numerous benefits can be gained if positive and negative matches are separated. This can be done using asymmetric similarity measures like Jaccard coefficient (Han & Kamber, 2006; Jaccard, 1901; Lourenco, Lobo, & Bacao, 2004). Unfortunately, the existing efforts for using Jaccard coefficient (Rajaraman et al., 2012) concentrate on finding the matches between two users irrespective whether they are positive or negative matches. This is contrary to the main objective of using Jaccard coefficient, i.e. emphasizing on positive matches between two users. For CRS, positive and negative matches give different meanings, positive matches are an indication for user satisfaction while negative matches indicate user dissatisfaction. If clearly specified, these two moods of the user can help CRS to direct its recommendation to the correct way.

In literature, Jaccard and Dice (Dice, 1945) coefficients are two coefficients that emphasize on the positive matches only and can be modified to take the negative matches into consideration. For Jaccard and Dice coefficients, the user profile consists of binary ratings, positive or negative ratings. Accordingly, four different subsets can be obtained for a pair of users, namely positive/positive ratings subset (P), positive/negative ratings subset (Q), negative/positive ratings subset (R), and negative/negative ratings subset (N) as illustrated in Table 1. Consequently, these four subsets constitute two main groups, namely agreement ratings group (P and N subsets) and disagreement ratings group (Q and R subsets).

As similarity measures, simple Jaccard and Dice coefficients pertain to asymmetric binary variables where negative matches have no information and therefore they are not suitable for CRS as they are. This motivates us to propose more representative variants of Jaccard and Dice coefficients which take into account the negative matches into consideration. This can be done by considering P and N subsets separately and hence a distinct similarity value is obtained for each subset. The final similarity value is obtained by aggregating positive and negative similarities using an aggregation function. A more general coefficient for asymmetric binary variable that takes the advantage of Jaccard and Dice coefficients is proposed in this paper. This coefficient is called a Power coefficient and it can include negative matches as discussed before for Jaccard and Dice coefficients. The main contributions of this paper are fourfold:

- Introducing many variants of Jaccard and Dice coefficients suitable for CRS.
- Proposing Power coefficient.
- Employing priority-based prediction formula.
- Introducing many scenarios for dividing a many point rating scale into bad and good subsets.

The rest of this paper is organized as follow: an introduction to some similarity measures for memory-based CRS is given in the next section. Jaccard coefficient for CRS is introduced in Section 3 while Section 4 presents Dice coefficient for CRS. Section 5 introduces the proposed Power coefficient for CRS. A priority-based prediction formula which keeps the predicted rating within the system's range of rating scale is introduced in Section 6. The experimental evaluation, methodology, experiments and results of the proposed approaches with the traditional approaches are presented in Section 7. Finally, we conclude our work in the last section.

Table 1
General classification of the ratings of a pair of users.

Rating 1	Rating 2	Subset	Group
Positive	Positive	P	Agreement
Negative	Negative	N	
Positive	Negative	Q	Disagreement
Negative	Positive	R	

2. Similarity measures for memory-based CRS in literature

Needless to say, personalized recommendations are the ultimate goal of any CRS. However, these recommendations depend extremely on the quality of the neighbors elected from the training set for the given active user. Whilst this set of neighbors cannot be selected correctly without a representative similarity measure.

Formally, CRS has M users, $U = \{u_1, \ldots, u_M\}$, having preferences for certain items such as products, news, Web pages, books, movies, restaurants, destinations, or CDs. The user's degree of preference for an item is represented by a rating that is obtained explicitly from the user directly or inferred implicitly from the users' navigation behavior. For example, if an Amazon customer views information about a product, the system can infer that he is interested in that product, even if he does not buy it. Each user, u_i rates a subset of items S_i from the K items, $S = \{s_1, \cdots, s_K\}$ of the system. The declared rating of user u_c for an item s_k is denoted by $r_{c,k}$ and the user's average rating is denoted by m_c. The set of users cross the set of items form a user-item matrix or a utility matrix (Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2013; Rajaraman et al., 2012; Schafer et al., 2007).

The set of ratings for each user forms the user profile which has to be compared to the profiles of other users based on a predefined similarity measure. The similarity between two users is a measure of how closely they resemble each other. Once similarity values are computed, the system ranks users according to their similarity values with the active user to extract a set of neighbors for him. According to the set of neighbors, the CRS assigns a predicted rating to all the items seen by the neighborhood set and not by the active user. The predicted rating, $pr_{x,k}$, indicates the expected interestingness of the item s_k to the user u_x.

The similarity between two users, u_x and u_y, based on Pearson correlation coefficient (Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2013; Rajaraman et al., 2012; Schafer et al., 2007) is computed only based on the common ratings, S_{xy}, both users have declared. The Pearson correlation coefficient (PCC) is:

$$corr(\mathbf{u}_x, \mathbf{u}_y) = \frac{\sum_{s_k \in S_{xy}} (r_{x,k} - m_x)(r_{y,k} - m_y)}{\sqrt{\sum_{s_k \in S_{xy}} (r_{x,k} - m_x)^2 \sum_{s_k \in S_{xy}} (r_{y,k} - m_y)^2}} \quad (1)$$

On the other hand, the cosine similarity measure (Adomavicius & Tuzhilin, 2005; Rajaraman et al., 2012) treats each user as a vector in the items' space and then takes the cosine of the angle between the two vectors as a similarity measure between the two users.

Download English Version:

https://daneshyari.com/en/article/382880

Download Persian Version:

https://daneshyari.com/article/382880

Daneshyari.com