Physica A 498 (2018) 102–115

Contents lists available at ScienceDirect

Physica A

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

Measuring transferring similarity via local information

Likang Yin [a,b], Yong Deng [a,c,*]

[a] Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China
[b] School of Hanhong, Southwest University, Chongqing 400715, China
[c] School of Computer and Information Science, Southwest University, Chongqing, 400715, China

HIGHLIGHTS

- This paper presents a proof of the range of classical transferring similarity.
- This paper proposes a novel method to fuse local similarity.
- We compare the proposed method with some common link prediction methods on 9 well-known datasets.

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form 22 November 2017
Available online 31 January 2018

Keywords:
Transferring similarity
Link prediction
Dempster–Shafer evidence theory
Belief function
Recommender systems

ABSTRACT

Recommender systems have developed along with the web science, and how to measure the similarity between users is crucial for processing collaborative filtering recommendation. Many efficient models have been proposed (i.g., the Pearson coefficient) to measure the direct correlation. However, the direct correlation measures are greatly affected by the sparsity of dataset. In other words, the direct correlation measures would present an inauthentic similarity if two users have a very few commonly selected objects. Transferring similarity overcomes this drawback by considering their common neighbors (i.e., the intermediates). Yet, the transferring similarity also has its drawback since it can only provide the interval of similarity. To break the limitations, we propose the Belief Transferring Similarity (BTS) model. The contributions of BTS model are: (1) BTS model addresses the issue of the sparsity of dataset by considering the high-order similarity. (2) BTS model transforms uncertain interval to a certain state based on fuzzy systems theory. (3) BTS model is able to combine the transferring similarity of different intermediates using information fusion method. Finally, we compare BTS models with nine different link prediction methods in nine different networks, and we also illustrate the convergence property and efficiency of the BTS model.

1. Introduction

Recent years have witnessed the exponential growth of the internet [1], the new challenge for modern science is the information overload. The personalized recommendations [2,3] is a efficient way to refine the massive data. Therefore, the recommender systems [4,5] come to a vogue in both the community of network science [6–9] and data mining field [10,11]. The recommender systems are a kind of complex entity that predicts the potential interesting by measuring the similarity between users' historical selections [12]. Generally, there are two classical approaches to implement the recommender

* Corresponding author at: Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China.

E-mail address: dengentropy@uestc.edu.cn (Y. Deng).

https://doi.org/10.1016/j.physa.2017.12.144
0378-4371/

systems [13]. The first one is the content-based method which makes the recommendation by finding the similar items according to the target user's historical selections [14]. The other one is called the user-based method [15,16]. Both of the two methods are based on the preferences of the users for a set of items [17]. Based on this inherent mechanism of recommender systems, how to properly measure the similarity between two users is of great significance. To address this issue, we can construct a bipartite network by regarding the users and items as different nodes and the accessing records as the links respectively [18,19]. Moreover, we can make commendations to users by predicting the missing links between unconnected user–item pairs in this bipartite network [20–22]. By this way, the issue of recommender systems has been transformed to the link prediction problems.

There are many efficient models in predicting the missing links in the networks. For example, the Common Neighbor (CN) index considers that the more common neighbors of two users represent the higher similarity they possess. Moreover, the Adamic–Adar (AA) index [23] and Resource Allocation (RA) [24] index have been proposed. AA and RA improve the accuracy by assigning a small-degree neighbor to a larger weight. Note that, RA and AA are called the similarity-based link prediction methods since both of them are based on the similarity between different nodes. There are some other kinds of prediction methods, such as the maximum likelihood methods [25], and the community-based models [26]. The link prediction is one of the hot topics in networks science, and there are some other valuable issues in network science, such as identifying the influential spreaders [27,28], supply chain networks [29], vulnerability analysis [30], network evolution [31,32], and self-similarity analysis [33].

By quantifying the similarity of users, some establishments and institutions have used the recommender systems in some practical applications, such as simulating spatial public goods games [34], analyzing the news consumption on Facebook [35], and video recommendation [36]. The similarity quantification has been widely used in both theoretical and practical fields, however, the similarity-based recommender systems still face challenges, such as the cold start problem [37], personalization problem [38,39], and measuring the effect of time [40,41].

However, how to handle the uncertainty of the high dimensional data is also an open issue in recommender systems [42–44]. For example, suppose there are three users, labeled as A, B, and C. There is a low similarity between A and C, but A and C are both very similar with B. In fact, the three users can share very similar potential preferences, and the low similarity between A and C may be caused by the sparsity of the data. That is to say, A and C share very few commonly selected items. The sparsity of dataset makes the direct similarity less accurate. To address this issue, the transferring similarity has been proposed [45]. However, the classical transferring similarity measure only presents an uncertain interval, and its range depends on the direct similarity degree. Therefore, the classical transferring similarity measure is hard to make accurate predictions.

The motivation of this study is to address both the issue of sparsity of datasets in the direct similarity measure and the issue of fuzzy similarity in classical models. The contributions of this paper are listed as follows. First, this paper proves the range of the interval of the classical transferring similarity. Second, the proposed method is able to fuse the transferring similarity of the intermediates (i.e., the common neighbors). Third, our model presents a specific state to denote the degree of similarity using fuzzy system theory. Moreover, to show the accuracy and efficiency, we compare the proposed method with nine common link prediction models on nine well-known datasets.

The structure of proposed method is listed as below. Given the raw user–item rating matrix, the first step is to calculate the direct similarity (i.e., the Pearson correlation) between two users. Second, based on the classical transferring similarity, we compute the upper bound and lower bound of the similarity interval. The third step is to construct the mass functions by discretizing the similarity interval. Next, we need to fuse the mass functions of all intermediates. Finally, we can make prediction according to the ultimate similarity state.

2. Preliminaries

2.1. Recommender systems

The recommender systems contain two main parts, one is the user and another is the object, and all users rated the each object in the target list. Hence, if there exists N users denotes as $U = u_1, u_2, \dots, u_N$ and M objects denotes as $O = o_1, o_2 \dots, o_M$, a rating $N \times M$ matrix can fully describe the input of the recommender systems. The common practice of recommender systems is to search for the preferences of a large group of people, and to find a small group of people who share the same taste as the target user. The recommender systems examine the preferences of these people, and combines them to create a ranking list of recommendations. There are various ways to help us determine who is close to our taste. Among them, there are two simple but very efficient methods, the first one is named Euclidean Distance Score (EDS) [46]. First, let features of the item, which are evaluated by different users, denote the multi-dimensional preference space, the coordinate of the user can be located according to the scores they give, and the distance between any two users reflects the degree of similarity between the two users. In general, the shorter the distance in the preference space, the higher the similarity between two users. Namely,

$$s_E(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{1}$$