Physica A 447 (2016) 199–207

Contents lists available at ScienceDirect

Physica A

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

Improving personalized link prediction by hybrid diffusion

Jin-Hu Liu [a], Yu-Xiao Zhu [a,*], Tao Zhou [a,b]

[a] Complex Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China

HIGHLIGHTS

- Personalized link prediction is introduced for the first time.
- Heat conduction process has been generalized to personalized link prediction.
- Two hybrid algorithms with great performance have been proposed.

ARTICLE INFO

Article history:
Received 24 July 2015
Received in revised form 23 October 2015
Available online 22 December 2015

Keywords:
Personalized link prediction
Heat conduction
Ground node

ABSTRACT

Inspired by traditional link prediction and to solve the problem of recommending friends in social networks, we introduce the personalized link prediction in this paper, in which each individual will get equal number of diversiform predictions. While the performances of many classical algorithms are not satisfactory under this framework, thus new algorithms are in urgent need. Motivated by previous researches in other fields, we generalize heat conduction process to the framework of personalized link prediction and find that this method outperforms many classical similarity-based algorithms, especially in the performance of diversity. In addition, we demonstrate that adding one ground node that is supposed to connect all the nodes in the system will greatly benefit the performance of heat conduction. Finally, better hybrid algorithms composed of local random walk and heat conduction have been proposed. Numerical results show that the hybrid algorithms can outperform other algorithms simultaneously in all four adopted metrics: AUC, precision, recall and hamming distance. In a word, this work may shed some light on the in-depth understanding of the effect of physical processes in personalized link prediction.

1. Introduction

Network has been used as a useful model to describe many social, biological and information systems, where nodes represent individuals and links reflect the relations or interactions between nodes [1–3]. Networks have been widely studied in many different fields and one of fundamental problems for network analysis is link prediction, which aims to estimate the likelihood of the existence of a link between two nodes based on observed links and the attributes of nodes [4,5]. For example, the existence of a link must be verified by costly chemical experiments in many biological networks, such as protein–protein interaction networks and metabolic networks. If the predictions are accurate enough, the experimental cost can be sharply reduced compared to blindly checking. Missing data problem also exists in social network, where link prediction is also one useful tool. In addition, link prediction algorithms can also be applied to identify spurious links [6–8]. Link prediction

* Corresponding author.
E-mail address: zhuyuxiao.mail@gmail.com (Y.-X. Zhu).

http://dx.doi.org/10.1016/j.physa.2015.12.036
0378-4371/

algorithms can not only be used to predict missing data but also practical to predict the links that may appear in the future of evolving networks. For example, in online social networks, very likely but not yet existent links can be recommended as promising friendships, which can help users to find new friends and thus enhance their loyalties to the websites.

In the traditional link prediction, all the nonexistent links are sorted in descending order according to their prediction scores, and the top-ranked links are most likely to exist. Clearly, in this case, the prediction list is generated from a global perspective, in which some nodes may have large number of promising links while others may have very few or even zero possible links. This straightforward and standard method may lead to some bias. On the one hand, in this case, the links that connect low-degree nodes may be ignored casually, while this kind of information may very be important and meaningful [9]. In addition, some research unveiled that low-degree users may have a big influence in the future [10]. On the other hand, the imbalance of prediction list may bring dissatisfaction for some individuals and thus affect the experience of the whole system. For example, in social networks, accurately predicting certain number of potential friends or acquaintances for each registered user is useful and meaningful. In this case, no real distinction can be made between low-degree and high-degree users and global link prediction does not apply in this case decently. However, this phenomenon has always been neglected in the traditional link prediction for the past several decades. To solve these problems, we propose personalized link prediction here, in which all nodes will get equal number of possible links through their own past link records.

One challenge deserved special attention recently, called low-diversity problem, has plagued almost all recommendation systems. It means that lots of recommender systems always recommend very similar items to different users which narrows users' views [11]. Subsequently, some physical dynamics, like heat conduction process (HC) have been applied to design recommender systems and can improve the diversity of recommendation. Motivated by this, we generalize heat conduction process to the framework of personalized link prediction and find that it outperforms other methods in diversity but does not perform very satisfactorily in accuracy. To solve this dilemma, ground node, that is supposed to connect all the nodes in the system, is incorporated to improve the prediction accuracy. Finally, we generalize one superior hybrid algorithm (LH) and propose another better hybrid algorithm (LGH) composed of local random walk (LRW) [12] and ground heat conduction (GHC), which performs pretty well not only on accuracy but also on diversity.

This article is organized as follows. In the next section, we will clearly define the problem of personalized link prediction, describe the standard metrics for evaluation. Then we explain several state-of-the-art similarity indices and introduce new algorithms HC, GHC, LH, LGH in Section 3. Data description and experimental results for the existed predicting algorithms and the proposed method are presented in Section 4. Finally, we summarize our results in Section 5.

2. Problem and metrics

For one given undirected network $G(V, E)$, in which V and E are the sets of nodes and links respectively. The universal set of all $\frac{|V|(|V|-1)}{2}$ possible links is denoted by U, where $|V|$ denotes the number of elements in set V (multiple links and self-connections are not allowed). Clearly the set of nonexistent links is $U \setminus E$, in which there are some missing links (i.e., the existed yet unknown links) and promising links (i.e., very likely but not yet existent links). The task of link prediction is to uncover these links. Each node pair x and y will be assigned a score s_{xy} according to a given prediction algorithm. The higher the score is, the higher the existence likelihood this link has. For each node x, we denote the set of its *relevant nonexistent links* (nonexistent links that connect x) as $(U \setminus E)_x$, thus all links in $(U \setminus E)_x$ are sorted in descending order according to their scores, and the top-ranked links are most likely to exist.

To test the performance of one given algorithm, we divide the observed links E into two sets: the training set E^T (considered as known information) and the test set E^P (used for testing and no information therein is allowed to be used for prediction). Clearly, $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. For each node x, the relevant test set (links in E^P that connect x) is denoted by E^P_x. We then introduce four popular evaluation metrics as below.

(i) AUC—short for area under the receiver operating characteristic curve, is considered as one standard metric to quantify the accuracy of prediction [13]. Specifically, for each node x, this metric can be interpreted as the probability that a randomly chosen relevant missing link (links in E^P_x) has higher score than a randomly chosen relevant nonexistent link (links in $(U\setminus E)_x$). In the implementation, among n times of independent comparisons, if there are n_1 times that the missing link has higher score and n_2 times the missing link and nonexistent link have the same score, the AUC value is defined as

$$\mathrm{AUC} = \frac{n_1 + 0.5n_2}{n}. \tag{1}$$

The AUC of the whole system is the average value over all nodes in the system. If all the scores are generated from an independent and identical distribution, the accuracy should be about 0.5. Therefore, the extent to which the accuracy exceeds 0.5 indicates how much better the algorithm performs than pure chance.

(ii) Precision and Recall [4]—Given the ranking of the non-observed links, the precision is defined as the ratio of relevant items selected to the number of items selected. Denoting by L the length of prediction list (i.e. the number of nodes recommended to each individual). For each individual x, if we take the top-L links as the predicted ones, among which L_x links are right (i.e., there are L_x links in the test set E^P_x), then the precision equals $P_x = L_x/L$. While recall is defined as the ratio of relevant items selected to the number of relevant items in the testing set. That is, $R_x = L_x/N_x$, where