Information Sciences 376 (2017) 158–171

Contents lists available at ScienceDirect

Information Sciences

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

Projection-based link prediction in a bipartite network

Man Gao [a], Ling Chen [a,b,*], Bin Li [a,b], Yun Li [a], Wei Liu [a], Yong-cheng Xu [a]

[a] Department of Computer Science, Yangzhou University, Yangzhou, 225009, China
[b] State Key Lab of Novel Software Tech, Nanjing University, Nanjing, 210093, China

ARTICLE INFO

Article history:
Received 6 December 2015
Revised 14 June 2016
Accepted 6 October 2016
Available online 11 October 2016

Keywords:
Bipartite network
Link prediction
Projection
Candidate node pairs

ABSTRACT

An algorithm for link prediction in a bipartite network is presented. In the algorithm, we first map the bipartite network onto a unipartite network called a projected graph. Based on the projected graph, we define the concept of a candidate node pair (CNP). We perform link prediction only within the CNPs to reduce the computation time. We also define the patterns covered by the CNPs and weights of the patterns. By calculating the weights of the patterns that a CNP covers, the connectivity of the CNP can be obtained, which can be used as the final score of link prediction. For a bipartite network with n and m nodes in the two parts, the time complexity of the proposed algorithm is $O(m)$, whereas those of other algorithms are $O(mn)$ or $O((m+n)^3)$. The experimental results show that our algorithm can achieve higher speed and superior quality link prediction results in bipartite networks compared with other methods.

1. Introduction

Networks can naturally describe various social structures. In such networks, vertices denote entities, and links represent communications or relations between the entities. A social network reflects persons or social organizations and their relations such as partnership or friendship. The analysis of social networking has drawn increasing attention in the field of sociology. It analyzes and explores the potential relations between social objects. In recent years, social network analysis has also attracted high interest in many business fields, such as e-business analysis and market modeling.

Link prediction is an important area in research on network analysis [41]. The objective of link prediction is to detect unobserved links from existing topologic features of the network or forecast future links from the current topologic structure of the network. In a social security network, link prediction is employed to discover underground groups of terrorists or criminals [28], whereas in networks of human behavior, link prediction is used to identify and categorize the activities and movement of people [1]. Link prediction also has many applications in networks reflecting social relations such as communication networks, email networks and sensor networks. In sensor networks, link prediction is used to discover dynamic temporal properties [44], to ensure information transfer secrecy [25], and to realize the most favorable routing [17].

Because relations among social members change continuously over time, links in real-world social networks are constantly varying and evolving. New links may appear, and existing links may vanish from the network. For example, email communications between friends, transactions between businesses, partnership between scientific researchers are changing over time. Therefore, link prediction algorithms should be capable of detecting dynamic relationships between members in

* Corresponding author. Fax: +8651487887937.
E-mail address: yzulchen@163.com (L. Chen).

http://dx.doi.org/10.1016/j.ins.2016.10.015
0020-0255/

a temporal social network. Recently, various approaches have been proposed to detect potential or future links in temporal social networks.

The similarity-based method is the most commonly used method for link prediction. In this method, each node pair is associated with an index to indicate the similarity between the corresponding nodes. This similarity quantifies the likelihood of link existence in the graph. Some essential attributes of the nodes can be used to define their similarity, such as the existence of many common features or topological structures between the nodes. Many studies in social networks show the existence of a relative similarity between individuals who are close to each other [2,11]. The structural similarity indices are often used in popular similarity-based methods. There are three types of similarity indices: local, quasi-local and global indices. Local indices are based on the neighbor information of the nodes, such as Common Neighbors, Jaccard, Salton, Sorensen, Preferential Attachment, Hub Depressed, Hub Promoted, Adamic-Adar, Resource Allocation, and Leicht–Holme–Newman (LHN1) indices [28]. Global indices require comprehensive information for link prediction tasks. They use global topological information of networks; such indices include Katz, Matrix Forest Index (MFI), and Leicht–Holme–Newman (LHN2) [28]. Quasi-local indices require less structural information than global indices but more information than local indices. Such indices include Superposed Random Walk (SRW) [27], Local Random Walk (LRW), and Local Path Index [48]. Compared with the local indices, global indices can achieve a more accurate solution of link prediction and require more computation time. Another class of similarity-based methods is random walk methods. Those methods include SimRank, Random Walk with Restart, Cos+, and Average Commute Time [28].

Some of these methods are based on the analysis of the topological features of the network. Purnamrita et al. [33] presented a nonparametric algorithm to predict potential links in dynamic networks. This algorithm partitioned the time domain into subsequences, which are represented by graph snapshots. Their method predicts connections between the nodes based on their topological features and local neighbors. For link prediction in social networks, they advanced a weighted-proximity-based method. Kim et al. [18] presented a method to predict future network topology using node centrality, which can identify important nodes in the future.

Machine learning strategies have also been exploited in network link prediction methods. Pujari et al. [32] applied a supervised rank aggregation method for link prediction in complex networks. Vu et al. [40] introduced a continuous-time regression model for temporal network link prediction. This model can be incorporated with both time-varying regression coefficients and time-dependent network statistics. Zeng et al. [45] presented a method using semi-supervised learning in link prediction utilize the latent information in the unlinked node pairs in networks. He et al. [15] proposed an algorithm for a link prediction ensemble based on an ordered weighted averaging operator. The algorithm assigns weights for nine local information-based link prediction algorithms and then aggregates their results to obtain the final prediction scores. Bao et al. [4] advanced a network link predictor using principal component analysis to identify features that are important to link prediction. Bringmann et al. [8] presented an algorithm for link prediction in networks based on techniques of association-rule mining and frequent-pattern detection. Using techniques for data mining and machine learning, the method can predict future co-participation of individuals in social events. To avoid high computational cost of optimization in the machine learning methods, some heuristic methods are employed in link prediction. Sherkat et al. [38] introduced an ant-colony-optimization-based algorithm for link prediction. Catherine et al. [7] proposed an approach to predict future links by applying the covariance matrix adaptation evolution strategy.

Some methods for link prediction on a network are based on probabilistic models. Hanneke et al. [14] proposed a family of statistical models for dynamic social network link prediction by extending the exponential random graph model. Liu et al. [26] presented a method for predicting potential links in a user-object bipartite network. Their method considers both the time attenuation and diversion delay. Gao et al. [12] advanced a model that exploits multiple information sources in the dynamic network to obtain link occurrence probabilities. Barbieri et al. [5] presented a stochastic link prediction model on directed graphs with node attribute features. In addition to predicting links, the model also provides explanations for the links detected. F. Hu et al. [16] presented a probabilistic model to detect human motion in social networks and advanced a method for labeling human motion using a constraint-based genetic algorithm to optimize the model. However, such a probabilistic model requires a predefined distribution of link appearance, which is difficult to know in advance for a given network.

The bipartite network is an important type of complex networks in real world applications, in which the nodes are partitioned into two parts such that no two nodes of the same part are adjacent. In a bipartite network, all edges connect only nodes in different parts, and no nodes of the same part are linked. Many real social networks are logically bipartite networks, such as the scientists–papers cooperation network [23,34], disease–gene network [10,47,20,19], RNA–protein interaction network [13], club members–activities network [36], term–document network [21], and investors–company network [39]. In recent years, link prediction in bipartite networks is applied in areas such as recommendation [9,31,42], social network analysis [6], and drug side effect prediction [29].

In recent years, many methods have been proposed to predict potential links in bipartite networks [6,9,22,24,29,43,37,46]. Supervised learning strategies have been exploited in bipartite network link prediction methods. Chang [9] et al. applied a supervised learning approach to predict potential links in the bipartite network representing Wikipedia. Benchettara et al. [6] introduce new topological features to quantify the probability of the link connecting the node pair in a bipartite network. Treating link prediction as a problem of binary classification, their approach uses a supervised machine learning approach to learn prediction models. However, those methods require a predefined distribution of link appearance, which is difficult to know in advance for a given network. By studying the structural holes in bipartite social networks, Xia et al. [43] presented