Information Sciences 329 (2016) 552–567

Contents lists available at ScienceDirect

Information Sciences

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

A fast algorithm for predicting links to nodes of interest

Bolun Chen [a,b,d], Ling Chen [a,c,*], Bin Li [a,c]

[a] Institute of Information Science and Technology, Yangzhou University, Yangzhou, China
[b] College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[c] National Key Lab of Novel Software Tech, Nanjing University, Nanjing, China
[d] Department of Physics, University of Fribourg, Chemin du Musee 3, Fribourg CH-1700, Switzerland

ARTICLE INFO

Article history:
Received 16 March 2015
Revised 13 August 2015
Accepted 26 September 2015
Available online 9 October 2015

Keywords:
Link prediction
Vertex similarity
Complex networks

ABSTRACT

The problem of link prediction has recently attracted considerable attention in various domains, such as sociology, anthropology, information science, and computer science. In many real world applications, we must predict similarity scores only between pairs of vertices in which users are interested, rather than predicting the scores of all pairs of vertices in the network. In this paper, we propose a fast similarity-based method to predict links related to nodes of interest. In the method, we first construct a sub-graph centered at the node of interest. By choosing the proper size for such a sub-graph, we can restrict the error of the estimated similarities within a given threshold. Because the similarity score is computed within a small sub-graph, the algorithm can greatly reduce computation time. The method is also extended to predict potential links in the whole network to achieve high process speed and accuracy. Experimental results on real networks demonstrate that our algorithm can obtain high accuracy results in less time than other methods can.

1. Introduction

Many social, biological, and information systems in the real world, from the nervous system to the ecosystem, from road traffic to the Internet, from an ant colony structure to human social relationships, can be naturally described as networks in which vertices represent entities and links denote relationships or interactions between vertices. As a topology approximation of complex systems, due to limitations of time and space, or experimental conditions, it is inevitable that there will be some errors or redundant links in constructing the complex network. At the same time, there will be some undetected potential links. In addition, because of the dynamic evolution of complex network links over time, we must predict missing and potential links according to known network information, which is the goal of the network link-prediction problem [27,32].

The link-prediction problem has a wide range of practical applications in various fields. For example, in biological networks, such as protein-protein interaction networks, metabolic networks and diseases-gene networks [22,43], links existing between nodes indicate that they have an interaction relationship. To mitigate the high costs of biological experimentation to reveal the hidden interaction relationships in these networks, the results of link prediction can direct biological experiments designed to reduce the cost and improve the success rate of the experiments. Predicting the loss and suspicious links of diseases-gene networks can help to explore the mechanisms of diseases, and predict and evaluate their treatment. Furthermore, it can also find new drug targets and open up new paths for drug development [12].

* Corresponding author at: Institute of Information Science and Technology, Yangzhou University, Yangzhou, China. Tel.: +86 514 87870026, fax: +86 514 87887937.
E-mail address: yzulchen@163.com, lchen@yzu.edu.cn (L. Chen).

http://dx.doi.org/10.1016/j.ins.2015.09.047
0020-0255/

In social network analysis, link prediction can also be used as a powerful supplementary tool to analyze accurately the social network structure. Studies on online social network analysis have been developing very rapidly in recent years. In online social networks, potential friendship of the users can be revealed by link prediction and can be recommended to the users [34]. By analyzing social relationships, we can find potential interpersonal links [7,9,16,20,21]. Link prediction can also be used in the academic network to predict the type and cooperators of an academic paper [38]. A link-prediction method can also be directly used for information recommendations [25,42] such as a commodity recommendation to customers [24,39]. Marketers would like to recommend products or services based on existing preferences or contacts. Social networking websites customize suggestions for new friends and groups using link prediction. For monitoring e-mail communication, link prediction is applied to detect anomalous e-mail [19]. Financial corporations must monitor transaction networks to detect fraudulent activity via link prediction. In monitoring networks of criminals, link prediction is used to discover hidden connections between criminals to prevent crime or terrorist activity.

Link prediction not only has a wide range of practical value but also has important theoretical significance. For example, link prediction is helpful to understand the mechanism of the evolution of a complex network [23,28,33]. Because the magnitude of the internal characteristics of a complex network structure is very large, it is difficult to compare the advantages and disadvantages of different mechanisms. Link prediction can provide a simple and unified platform for a fair comparison of network evolution mechanisms to promote theoretical research on the complex-network evolution model.

In many real world network applications, we must detect the most-possible links connecting with a given vertex in the network. We must answer queries such as, "Which are the K most-possible links connecting with vertex v in the network?"; "Which five authors share the most-similar research interests with Professor Johnson?"; "Who are the ten customers with the most-similar shopping habits with customer John Smith?"; "What are the closest ten proteins to a given myoglobin?"; and so on. These are actually link-prediction problems on a given vertex. In fact, in many real world applications, we must predict similarity scores only between pairs of vertexes that users are interested in rather than predicting the scores of all pairs of vertexes in the network.

To answer precisely such a link prediction query for a given vertex v using global indices, it is not possible to independently calculate the indices of links connecting with v. Due to the global nature of global indices calculations, we still must calculate the indices of all the node pairs in the network, although we are only interested in the ones involving v. However, calculating the indices of all the node pairs in the network requires a large amount of computation time.

In this paper, we propose a fast similarity-based method to predict links related to the nodes in which users are interested. The method first constructs a sub-graph centered at the node of interest. For a given error bound ε, we can choose the size of such a sub-graph to make the error of the estimated similarities be less than ε. Because the algorithm computes similarity scores only within a small sub-graph, the computation time is greatly reduced. The method is also extended to predict potential links in the whole network and to achieve high process speed and accuracy. Our experiment results on real networks show that the algorithm can achieve higher speed and more accurate results than can other methods.

The rest of this paper is organized as follows. Section 2 reviews related work on link prediction in complex networks. Section 3 reviews methods based on local random walk. Section 4 defines the r-radius sub-graph of a node v in network G for a given error bound ε. Section 5 presents the fast r-radius sub-graph-based algorithm *Single_Node-LP* to predict the links related to the nodes in which the users are interested, and the sub-graph-based algorithm *Node-LP* to predict the links in the whole network. Section 6 shows and analyzes the experimental results obtained by the algorithms *Single_Node-LP* and *Node_LP* and compares their performance with other similar methods. Section 7 presents the conclusions.

2. Related works

In recent years, many methods of link prediction have been reported. Those methods can be classified into three categories: similarity-based methods, machine-learning methods and probabilistic model-based methods.

The similarity-based method is the most commonly used method for link prediction. In the similarity-based method, each node pair is assigned an index, which is defined as the similarity between the two nodes. All non-observed links are ranked according to their similarities, and the non-observed links connecting nodes that are more similar are supposed to have higher existence likelihoods. Node similarity can be defined by using the essential attributes of nodes: two nodes are considered similar if they have many common features or correlated topological structures [1,15,26]. Many studies found that there are substantial levels of topical similarity among individuals who are close to one another in the social network. For instance, Aiello et al. [2] studied friendship prediction in social networks based on the presence of homology in three systems that combine tagging social media with online social networks. Many works exploit topological features of network structures for link-prediction tasks. S. Gao et al. [11] defined the overall relationships between object pairs as a link pattern, which consists of an interaction pattern and a connection structure in the network. The structural similarity indices can be classified into three categories: local indices, global indices, and quasi-local indices. Local indices use only neighbor information of the nodes. Typical local indices include Common Neighbors, the Salton Index, the Jaccard Index, the Sorensen Index, the Hub Depressed Index, the Hub Promoted Index, the Leicht–Holme–Newman Index, the Preferential Attachment Index, the Adamic–Adar Index and the Resource Allocation Index [32]. Global indices require global topological information. The Katz Index, the Leicht–Holme–Newman Index and the Matrix Forest Index [32] are typical global indices. Quasi-local indices do not require global topological information but make use of more information than do local indices. Such indices include the Local Path Index [31,45], Local Random Walk, and Superposed Random Walk [30]. Another similar group is based on the random walk. These include indices such as Average Commute Time, Cos+, Random Walk