Physica A 436 (2015) 216–223

Contents lists available at ScienceDirect

Physica A

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

Predicting missing links via correlation between nodes

Hao Liao [a,b], An Zeng [a,d,*], Yi-Cheng Zhang [c,e]

[a] Alibaba Research Center for Complexity Sciences, Alibaba Business College, Hangzhou Normal University, Hangzhou 311121, PR China

[b] Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, PR China

[c] Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

[d] School of Systems Science, Beijing Normal University, Beijing 100875, PR China

[e] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China

HIGHLIGHTS

- We introduce a new link prediction method based on the Pearson correlation.
- The new method is very effective in calculating similarity based on high order paths.
- We combined the correlation method with the resource allocation method.
- The combined method is effective in solving the cold-start and data sparsity problem.

ARTICLE INFO

Article history:
Received 2 November 2014
Received in revised form 9 February 2015
Available online 11 May 2015

Keywords:
Link prediction
Correlation coefficient
Node similarity

ABSTRACT

As a fundamental problem in many different fields, link prediction aims to estimate the likelihood of an existing link between two nodes based on the observed information. Since this problem is related to many applications ranging from uncovering missing data to predicting the evolution of networks, link prediction has been intensively investigated recently and many methods have been proposed so far. The essential challenge of link prediction is to estimate the similarity between nodes. Most of the existing methods are based on the common neighbor index and its variants. In this paper, we propose to calculate the similarity between nodes by the Pearson correlation coefficient. This method is found to be very effective when applied to calculate similarity based on high order paths. We finally fuse the correlation-based method with the resource allocation method, and find that the combined method can substantially outperform the existing methods, especially in sparse networks.

1. Introduction

The ultimate objective of many scientific studies is to do prediction. For instance, understanding the mechanism of epidemic spreading can help us to predict the future coverage of a certain virus [1], the mechanistic model for the citation dynamics of individual papers can be applied to predict the future evolution of scientific publications [2]. While mathematical models and prediction techniques are sufficiently mature for some systems, reliable prediction approaches are still unavailable in most systems. Besides the prediction of the collective behavior, the prediction in microscopic level, such as the well-known link prediction challenge in complex networks, has also attracted a lot of attention.

* Corresponding author at: School of Systems Science, Beijing Normal University, Beijing 100875, PR China.
E-mail address: anzeng@bnu.edu.cn (A. Zeng).

http://dx.doi.org/10.1016/j.physa.2015.05.009
0378-4371/

Link prediction is a very important problem that aims at estimating the likelihood of the existence of a link between two nodes [3]. Solving this problem cannot only help us complete the missing data in biological networks such as the protein–protein interaction networks and metabolic networks [4,5], but also enable us to predict the evolution of social networks [6,7]. In fact, link prediction is also closely connected to some other problems such as recommendation [8] and spurious links detection [9]. A sound link prediction method will help to design more efficient recommendation algorithm to filter out irrelevant information for online users [10]. Moreover, the link prediction method can also be applied to analyzing the reliability of existing links and accordingly identifying some noisy connections of networks. The progress in this field will largely push forward the research in other fields. Accordingly, the problem of missing link prediction has been intensively studied by researchers from different backgrounds and many methods applied to different fields have been proposed [11–14]. For a review, see Ref. [15].

The basic assumption for link prediction is that two nodes are more likely to have a link if they are similar to each other. Therefore, the essential problem for link prediction is how to calculate the similarity between nodes accurately. One of the most straightforward methods is called common neighbor which measures the similarity between two individuals by directly counting the number of common neighboring nodes [16]. However, this method has serious shortcomings as it strongly favors the large degree nodes. To solve this problem, many variants, such as the Jaccard index [17] and Salton index [18], have been applied to remove this tendency. In addition, some other methods including Katz index [19], simrank [20], hierarchical random graph [5] and stochastic block model [21,22] are also very effective in estimating nodes' similarity. However, these methods are based on global algorithms that can be prohibitive to use for large-scale systems.

In this paper, we argue that the similarity between nodes can be calculated based on another completely different type of method, namely correlation coefficient. In broad definition, correlation refers to any class of statistical relationships involving dependence between two or more random variables. In our case, it actually refers to the Pearson correlation [23] between nodes' attribute vectors which can come from the adjacency matrix or higher order of that. In link prediction, one of the biggest challenges is the data sparsity. It means that a lot of data at hand is too sparse to extract valuable similarity information from the simple common neighbor method or its variants. One possible solution has been discussed in Ref. [24] in which longer paths (i.e. paths with length larger than 2) are applied to measure nodes' similarity. However, when it comes to such high order information, much noise will be included so that the similarity matrix is indeed denser but the similarities are not satisfactorily accurate, which leads to a poor outcome of predicted links. In our simulation, we find that the correlation-based method is very effective when applied to calculating similarity based on high order paths. We finally use the new method with the resource allocation method [25], and find that the combined method can substantially outperform the existing methods, especially in sparse networks.

2. Related works

To begin our analysis, we first briefly describe the link prediction problem and review some representative methods. Considering an unweighted undirected simple network $G(V, E)$, V is the set of nodes and E is the set of links. The multiple links and self-connections are not allowed. For each pair of nodes x, y belonging to V, we calculate a score s_{xy} which measures the likelihood for nodes x and y to have a link between them. Since G is undirected, the score is supposed to be symmetry, i.e. $s_{xy} = s_{yx}$. All the nonexistent links are sorted in a decreasing order according to their s scores, and the links on the top are more likely to exist. There are many different ways to calculate s_{xy} score and the most common and straightforward way is to calculate the similarity between nodes x and y.

Generally speaking, two nodes are considered to be similar if they have some common important features in topology [15]. In this paper, we compare the prediction accuracies of four typical similarity indices: Common neighbor (CN), Resource allocation (RA), Jaccard and Local path. Their definitions and relevant motivations are introduced as follows:

(i) *Common Neighbor* (CN). Two nodes x and y are more likely to form a link if they have many common neighbors. Let $\Gamma(x)$ denote the set of neighbors of node x, the simplest measure of the neighborhood overlap can be directly calculated as:

$$s_{xy}^{\mathrm{CN}} = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

which is the actual aggregation method used by most websites. However, the drawback of CN is that it is in favor of the nodes with large degree. It is obvious that $s_{xy} = (A^2)_{xy}$, where A is the adjacency matrix. $A_{xy} = 1$ if x and y are directly connected and $A_{xy} = 0$ otherwise. Note that $(A^2)_{xy}$ is also the number of different paths with length 2 connecting x and y. Newman [15] used this quantity in the study of collaboration networks, showing the correlation between the number of common neighbors and the probability that two scientists will collaborate in the future. Therefore, we here select CN as the representative of all CN-based measures. Although CN consumes little time and performs relatively good among many local indices, due to the insufficient information, its accuracy cannot catch up with the measures based on global information. One typical example is the Katz index [19].

(ii) *Jaccard coefficient* (Jaccard). This index was proposed by Jaccard over a hundred years ago. The algorithm is a traditional similarity measurement in the literature. It is defined as

$$s_{xy}^{\mathrm{Jaccard}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}. \tag{2}$$