Knowledge-Based Systems 89 (2015) 669–680

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Node-coupling clustering approaches for link prediction

Fenhua Li [a,d,e,*], Jing He [b], Guangyan Huang [f], Yanchun Zhang [b,d,*], Yong Shi [a,c], Rui Zhou [b]

[a] *CAS Research Center on Fictitious Economy and Data Science, University of Chinese Academy of Sciences, Beijing 100190, China*
[b] *Centre for Applied Informatics, Victoria University, Melbourne, Australia*
[c] *CAS Key Laboratory for Big Data Mining and Knowledge Management, Beijing, China*
[d] *School of Computer Science, Fudan University, Shanghai, China*
[e] *Department of Computer Science and Technology, Yuncheng University, Shanxi, China*
[f] *School of Information Technology, Deakin University, Melbourne, Australia*

ARTICLE INFO

Article history:
Received 4 March 2015
Received in revised form 10 July 2015
Accepted 10 September 2015
Available online 15 September 2015

Keywords:
Link prediction
Node-coupling clustering
Data mining
Big data

ABSTRACT

Due to the potential important information in real world networks, link prediction has become an interesting focus of different branches of science. Nevertheless, in "big data" era, link prediction faces significant challenges, such as how to predict the massive data efficiently and accurately. In this paper, we propose two novel node-coupling clustering approaches and their extensions for link prediction, which combine the coupling degrees of the common neighbor nodes of a predicted node-pair with cluster geometries of nodes. We then present an experimental evaluation to compare the prediction accuracy and effectiveness between our approaches and the representative existing methods on two synthetic datasets and six real world datasets. The experimental results show our approaches outperform the existing methods.

1. Introduction

1.1. Background

With the rapid development of internet technology, the amount of information in social networks increases significantly. While accessing useful information from social networks has become more and more difficult [1]. Social networks contain large number of potential useful information that is valuable for people's daily lives and social business [2]. Therefore, social network analysis (SNA) has become a research focus to mine latent useful information from massive social network data. As part of this research, how to accurately predict a potential link in a real network is an important and challenging problem in many domains, such as recommender systems, decision making and criminal investigations. For example, we can predict a potential relationship between two persons to recommend new relationships in the Facebook network. In general, we call the above problem as link prediction [3].

As a subset of link mining [4], link prediction aims to compute the existence probabilities of the missing or future links among vertices in a network [5,6]. There are two main difficulties in the link prediction problem: (1) huge amount of data, which requires the prediction approaches to have low complexity and (2) prediction accuracy, which requires the prediction approaches to have high prediction accuracy. However, traditional data mining approaches cannot solve the link prediction problem well because they do not consider the relationships between entities, but the links between entities in a social network are interrelated.

To overcome the above two difficulties and meet the practical requirements, many similarity-based methods have been proposed. These methods are mainly based on local analysis and global analysis [7]. The approaches based on local analysis consider only the number or different roles of the common neighbor nodes, which results in lower time complexity. At the same time, they have lower accuracy because of insufficient information. On the other hand, the approaches based on global analysis have higher prediction accuracy and higher time complexity due to accessing the global structure information of a network [5,8]. So these methods are not satisfying solutions that can overcome the aforementioned two difficulties.

In this paper, we propose two novel node-coupling clustering approaches and their extensions for the link prediction problem. They consider the different roles of nodes, and combine the coupling degrees of the common neighbor nodes of a predicted

* Corresponding authors at: CAS Research Center on Fictitious Economy and Data Science, University of Chinese Academy of Sciences, Beijing 100190, China (F. Li). Centre for Applied Informatics, Victoria University, Melbourne, Australia (Y. Zhang). Tel.: +86 18810401728.

E-mail addresses: llifenhua@126.com (F. Li), Yanchun.Zhang@vu.edu.au (Y. Zhang).

http://dx.doi.org/10.1016/j.knosys.2015.09.014
0950-7051/© 2015 Elsevier B.V. All rights reserved.

node-pair with cluster geometries of nodes. Our approaches remarkably outperform the existing methods in terms of efficiency accuracy and effectiveness. This is confirmed by experiments in Section 5.

1.2. Contributions

The contributions of this paper consist of the following three aspects: (1) We propose two novel node-coupling clustering approaches and their extensions, which define a novel node-coupling degree metric. (2) We consider the coupling degrees of the common neighbor nodes of a predicted node-pair, by which some links that the existing methods cannot predict are accurately predicted. (3) We use the clustering coefficient to capture the clustering information of a network, which makes our approaches have lower time complexity compared with the existing clustering methods. (4) We use the clustering information that is important information for predicting links, which can improve the prediction accuracy. Experimental evaluation demonstrates our approaches outperform other methods in terms of accuracy and complexity. Our approaches are very suitable for large-scale sparse networks.

1.3. Organization

The rest of this paper is organized as follows: Section 2 provides the overview of the related works of link prediction. Some preliminaries are briefly introduced in Section 3. Section 4 presents the idea of our approaches, and gives their complexity analysis. Experimental study is presented in Section 5. Section 6 concludes this paper and the future work.

2. Related work

The existing link prediction approaches can be divided into three categories: the methods based on local analysis and global analysis [7], maximum likelihood estimation methods [5], and machine learning methods [5].

The methods based on local analysis and global analysis exploit the similarity of nodes in a network. The methods based on local analysis consist of Common Neighbors (*CN*), Adamic Adar (*AA*), Preferential Attachment (*PA*) and Jaccard Coefficient (*JC*). They suppose that the nodes of a network are independent of each other, and perform mostly on the local structure information of a network (i.e. node degree, nearest neighbors information). In contrast, there are Katz (*Katz*), Hitting Time, Average Commute Time (*ACT*), Cosine based on random walk (*CosRW*), Graph distance (*GD*), Rooted PageRank and SimRank in the methods based on global analysis. These methods capture the global structure information of a network (i.e. the path set of a specific length). These above methods are described and discussed in [5,7]. Besides, Liu and Lü introduced a local random walk approach that provides good prediction accuracy [9]. Furthermore, Zhou et al. proposed two similarity-based methods – Resource Allocation (*RA*) and Local Path (*LP*). They verified the performance of these two methods on six real-world datasets [10]. In practice, since the methods based on local analysis focus only on the local structure information of a network, they have lower computational complexity than those based on global analysis, and are suitable for large-scale networks. However, as there are no sufficient information in the local-based methods, these methods have lower prediction accuracy than those based on global analysis.

Maximum likelihood estimation methods apply some presumed rules and parameters with the maximum probability of the known structure to predict the potential links in a network. Clauset et al. proposed a missing-links prediction method based on hierarchical structure in the incomplete networks [11]. Guimera and Sales-Pardo presented a stochastic block model approach in link prediction [12].

Machine learning methods use the learned model to predict links by extracting the latent structure information of a known network. O'Madadhain et al. applied several classifiers to predict the potential links from probability theory in the networks based on events [13]. Hasan et al. treated link prediction as a supervised learning task, which predicts the link of a predicted node-pair by identifying a negative or positive example. They extracted the features of a co-authorship network and evaluated the prediction results of several classifiers [14].

Although the above latter two groups methods provide competitively accurate prediction compared with the methods based on local analysis and global analysis, they are only suitable for small scale real networks and impractical for large-scale sparse networks because of their high computational complexities. Therefore, Li and He et al. presented a clustering-based link prediction method (i.e. *CLPA*). They considered the clustering and free scale feature of a network for link prediction in their method [15]. In this paper, we will propose two novel node-coupling clustering approaches and their extensions to solve this problem.

3. Preliminaries

3.1. Clustering coefficient

In graph theory, clustering coefficient is a metric that can evaluate the extent to which nodes tend to cluster together in a graph [16]. It can capture the clustering information of nodes in a graph [17]. An undirected network can be described as a graph $G = (V, E)$, where V denotes the set of nodes and E indicates the set of edges. $v_i \in V$ is a node in Graph G. The clustering coefficient of node v_i in Graph G can be defined as

$$C(i) = \frac{E_i}{(k_i \cdot (k_i - 1))/2} = \frac{2 \cdot E_i}{(k_i \cdot (k_i - 1))} \qquad (1)$$

where $C(i)$ denotes the clustering coefficient value of node v_i. k_i represents the degree value of node v_i. E_i is the number of the connected links among k_i neighbors of node v_i. For example, there is a node v_1 in Graph G. The degree value of node v_1 is 5 (i.e. $k_1 = 5$). The number of connected links among the neighbors of node v_1 is 6 (i.e. $E_1 = 6$). Thus, the clustering coefficient value of node v_1 is: $C(1) = \frac{2 \cdot E_1}{(k_1 \cdot (k_1 - 1))} = \frac{2 \cdot 6}{5 \cdot (5-1)} = 0.6$.

3.2. Evaluation metrics

In this section, we present two popular metrics for link prediction accuracy employed in this paper – Area under the receiver operating characteristic curve (*AUC*) and *Precision*. In general, a link prediction method can compute a score S_{xy} for each unknown link to evaluate its existence probability and give an ordered list of all unknown links based on these S_{xy} values [18].

3.2.1. AUC

It can evaluate the overall performance of a link prediction method. As described in [5,8], the *AUC* value can be considered as the probability that the S_{xy} value of an existing yet unknown link is more than that of a non-existing link at random. That is, we randomly select an existing yet unknown link in the test set and compare its score with that of a non-existing link at a time. There are N independent comparisons, where the times that the existing yet unknown links have higher S_{xy} value are H, and the times that they have the same S_{xy} value are E. The *AUC* value is defined as: