Journal of Computational Science 13 (2016) 1–9

Contents lists available at ScienceDirect

Journal of Computational Science

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

Adaptive degree penalization for link prediction

Víctor Martínez*, Fernando Berzal, Juan-Carlos Cubero

CITIC & Department of Computer Science and Artificial Intelligence, University of Granada, Spain

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 12 June 2015
Accepted 10 December 2015
Available online 24 December 2015

Keywords:
Link prediction
Networks
Graphs
Topology
Shared neighbors

ABSTRACT

Many systems of interest are best described using networks that represent binary relationships among their elements. Link prediction aims to infer the link formation process by predicting missed or future relationships based on currently observed connections. Different techniques and measures have been proposed in the literature to solve this problem. Similarity-based local methods achieve high precision with a low computational complexity. However, determining which particular technique should be applied for each particular network remains an open question. In this paper, we exploit the existence of a relationship between the best-performing degree of penalization for shared neighbors and the network clustering coefficient. We propose an adaptive degree penalization link prediction method, a novel link prediction technique that achieves better results than previously proposed methods.

1. Introduction

The link prediction problem consists of inferring the formation of new relationships or the existence of still-unknown connections between pairs of entities in a network based on their properties and currently observed links [24]. This problem has attracted a lot of attention, since a large number of systems in many different fields can be described using networks. Approaches and techniques to solve this problem allow us to extract implicit information present in the network, identify spurious links, or model and evaluate network evolution mechanisms. These problems are of great interest since they are closely related to other problems usually found in different disciplines. For example, link prediction has been used to predict previously unknown protein interactions in protein-protein interaction networks [27]. It has also been used to study and predict future author collaborations and tendencies in co-authorship networks [34]. In fact, link prediction is present in our daily lives when we get friendship suggestions in social networks [10] or recommendations of new products in e-commerce web sites [16].

The link prediction problem is formally defined as follows. Let G be an undirected graph $G=(V, E)$, where V is a set of optionally labeled nodes and E is a set of edges (also referred to as links) between pairs of elements from set V. Given a snapshot of the network G at time t, the link prediction problem consists of inferring the subset of missing links in the current snapshot that will be formed at time $t+\Delta$.

Some notational conventions are important to properly describe the proposed solutions for the link prediction problem. An edge between nodes x and y is denoted as $e_{x,y}$. The number of nodes in the network is $|V|$. The number of edges is $|E|$. The set of nodes connected through an edge to a node x is called the neighborhood of x and is referred to as Γ_x. The degree of a node x in an undirected graph is defined as the number of edges connected to the node and will be denoted as $|\Gamma_x|$.

Many link prediction methods are based on the observation that nodes that share a higher number of neighbors are more likely to be connected [30]. Well-known link prediction techniques take into account the number of directly shared neighbors (local methods) or the number of chains of neighbors between two nodes (global methods) to estimate the probability of the existence of a potential link. However, these techniques always work the same way, regardless of the network they are applied to. Our work is motivated by the lack of further studies about how link prediction techniques are affected by network structural properties and how existing methods can be adapted to the structural properties of particular networks in order to obtain better results.

This paper is organized as follows. Related work is presented in Section 2. We propose and describe a generalized degree penalization similarity measure in Section 3. In Section 4, we analyze the relationship of the best-performing degree penalization with respect to the topological properties of the network. A novel link prediction technique called adaptive degree penalization is presented in Section 5. Finally, the conclusions drawn from this study and some lines of future research are presented in Section 6.

* Corresponding author.
E-mail address: victormg@acm.org (V. Martínez).

http://dx.doi.org/10.1016/j.jocs.2015.12.003
1877-7503/

2. Related work

Link prediction has been the subject of many studies [12]. A large number of techniques following different approaches have been proposed to deal with the link prediction problem [26]. In this work, we limit our scope to techniques that consider only network topology, albeit methods considering other attributes have also been proposed [13].

The first and most studied approach is based on the similarity between nodes [24,26]. Similarity-based techniques assume that nodes are more likely to form links with similar nodes. A function that assigns a similarity score $s(x, y)$ to every pair of nodes in the network is defined. This similarity score can take into account different features, which can be topological properties or network-specific attributes. All possible pairs of nodes are ranked in decreasing order based on their similarity scores. Links at the top of the ranked list are supposed to be more likely to be present in the set of missing links.

Similarity-based methods can be categorized depending on the amount of information taken into consideration when computing the similarity function. For example, local similarity techniques consider only direct neighbor information. This family of techniques can achieve high precision in most networks and have a linear time complexity, which makes them suitable for large networks. On the other hand, global methods use the whole topology of the network to compute the similarity score for every possible link. This type of techniques has the advantage of being able to compute the similarity between each pair of nodes regardless of their distance within the network, instead of being limited to neighbor-sharing pairs of nodes. Their main drawbacks are their high computational complexity and their sensitivity to noise, which usually leads to lower precision than local methods. Finally, quasi-local techniques have been proposed to try to find an equilibrium between the amount of considered information and the computational complexity of the resulting methods. Most quasi-local techniques are either based on local ones with small variations to consider neighbors of neighbors or based on global ones with constraints on the lengths of the considered paths.

An alternative approach is to describe the network formation model in statistical terms. Statistical approaches build a parameterized model assuming the existence of a known structure in the network [15,42,7,14]. The parameters of the model for a particular network are estimated using statistical methods. Finally, the adjusted model is used to compute the probability of the formation of each possible link. The main problem of this kind of techniques is that they suffer from a very high computational cost, which limits their applicability to networks of only hundreds or a few thousand nodes. In addition, they can only be applied to networks with a particular structure.

Other algorithmic approaches have also been proposed. Since link prediction techniques are inherently heuristic, some metaheuristic-based methods have been proposed in order to automatically adjust the influence of a set of local similarity-based techniques in an attempt to maximize precision [5]. The link prediction problem can be seen as a classification problem with two classes (existence and absence of links). This point of view allows the application of traditional machine learning techniques [19,8]. These techniques can obtain better results than other approaches at the cost of a previous training stage, which is not always possible in many applications. Furthermore, they have the drawback that the predictive model they build is often hard to understand and analyze.

In this paper, we focus our attention on local similarity-based techniques, since these techniques are widely used due to their high scalability and the reasonable precision they obtain [45]. Computational complexity is really important in link prediction, since most real networks are huge, with hundreds of thousands or even millions of nodes. Even worse, there are usually time or resource constraints in many problems related to link prediction. In fact, most recommender systems use only local techniques.

The most basic local method is called Common Neighbors (CN). This technique assigns a score based just on the number of shared neighbors:

$$s^{CN}(x, y) = |\Gamma_x \cap \Gamma_y| \tag{1}$$

It makes sense to assume that, if two individuals share many acquaintances, they are more likely to meet than two individuals without common contacts. Different studies have confirmed this hypothesis by observing a correlation between the number of shared neighbors between pairs of nodes and their probability of being linked [30].

Lada Adamic and Eytan Adar proposed the Adamic-Adar Index (AA) to measure the similarity between two entities based on their shared features [1]. This measure was adapted to link prediction by considering shared neighbors as features:

$$s^{AA}(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log |\Gamma_z|} \tag{2}$$

This equation is a variation of the common neighbors similarity function. Here, each shared neighbor is penalized by its degree. This intuitively makes sense in a large number of real-world networks. For example, in social networks, the amount of resources or time that a node can spend on each of its neighbors decreases as its degree increases, also decreasing its influence on them.

The Resource Allocation Index (RA) was motivated by the resource allocation process which takes place in complex distribution networks [45]. It models the transmission of resources between two unconnected nodes x and y through neighborhood nodes. Each neighborhood node gets a given amount of resources and distributes them evenly among its neighbors. The amount of resources obtained from node x by node y through their shared neighbors can be considered as a similarity measure between both nodes. The resource allocation index has shown to be the local measure with better results in a large number of networks [25]. It can be computed as

$$s^{RA}(x, y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{|\Gamma_z|} \tag{3}$$

Other local similarity-based techniques have been proposed, including the Preferential Attachment Index [3], the Jaccard Index [18], the Salton Index [37], the Sørensen Index [40], the Hub-Promoted and Hub-Depressed Indices [36], and the Leicht-Holme-Newman Index [22]. Most of these techniques are variations of the previously described measures, but also consider other features such as the number of unshared neighbors. Different comparative studies have shown that these variations work better in very specific contexts, yet are worse on average [45].

3. Similarity based on adjustable degree penalization

The Common Neighbors method, the Adamic-Adar Index, and the Resource Allocation Index have been presented in the literature as three different link prediction techniques. It can be readily seen that these methods assume that the probability of existence of a link between two nodes is proportional to the number of shared neighbors between them, but penalize each one according to their degree, with a null penalization in the Common Neighbors case. From our point of view, CN, AA, and RA are just variations of the same technique, which considers shared neighbors and penalizes them by their degree using different penalization schemes.