Ciência e Natura, v. 37 Part 2 jun. 2015, p. 125–132

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

ciência e natura

Link Prediction in Social Networks Using Markov Random Field

Zohreh Zalaghi

Computer Group, Facultu of Engineering, Islamic Azad University, Doroud Branch

Abstract

Link prediction is an important task for social networks analysis, which also has applications in other domains such as information retrieval, recommender systems and e-commerce. The task is related to predicting the probable connection between two nodes in the netwok. These links are subjected to loss because of the improper creation or the lack of reflection of links in the networks; so it`s possible to develop or complete these networks and recycle the lost items and information through link prediction. In order to discover and predict these links we need the information of the nodes in the network. The information are usually extracted from the network`s graph and utilized as factors for recognition. There exist a variety of techniques for link prediction, amongst them, the most practical and current one is supervised learning based approach. In this approach, the link prediction is considered as binary classifier that each pair of nodes can be 0 or 1. The value of 0 indicates no connection between nodes and 1 means that there is a connection between them. In this research, while studying probabilistic graphical models, we use Markov random field (MRF) for link prediction problem in social networks. Experimentl results on Flicker dataset showed the proposed method was better than previous methods in precision and recall.

Keywords: *Social Network, Link Prediction, Supervised Learning, Probabilistic Graphical Model, MRF.*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

The analysis of social networks has recently experienced a surge of interest by researchers, due to different factors, such as the popularity of online social networks (OSNs), their representation and analysis as graphs, the availability of large volumes of OSN log data, and commercial/marketing interests [1]. As the size and number of online social networks are increasing day by day, social network analysis has become a popular issue in many branches of science. One of the emerging topics in social network analysis is link prediction. Prediction of a new connection or link between two nodes based on attributes of existing nodes and links in the graph is called link prediction [2]. Link prediction can be categorized into two classes [3]: (1) Problem of identifying existing yet unknown links; (2) Predicting links that may appear in the future.

The social network is represented as a graphic structure made up of a set of nodes and links, where nodes represent the individuals within network and links denote the relationships between individuals [4]. People use social networks to communicate, collaborate, and share information. One of the most profound properties of social networks is their dynamic nature. People join and leave social networks. Also, the circle of friends may frequently change when people establish friendship through social links or when their interest in a social relationship ends and the link is removed [5]. In this paper, we propose a link prediction approach using Markov random field (MRF). The rest of this paper is organized as follows: Section 2 provides the related work on the context of social networks and link prediction. In section 3 proposed method is described. Experiments and results analysis are given in section 4. Finally in Section 5, the conclusions are discussed.

2 Related Work

In recent years, many methods on link prediction have been reported. Those methods can be classified into categories such as similarity-based methods, maximum likelihood methods and probabilistic model based methods.

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 links connecting more similar nodes are supposed to have higher existence likelihoods[6]. Many studies found that there are substantial levels of topical similarity among users who are close to each other in the social network, such as friendship prediction in [7], which studied the presence of homology in three systems that combine tagging social media with online social networks.

Another category of link prediction method is based on maximum likelihood estimation. These methods presuppose some organizing models of the network structure, with the detailed rules and specific parameters obtained by maximizing the likelihood of the observed structure. Then, the likelihood of any non-observed link can be calculated according to those rules and parameters. Typical organizing models of the network are the hierarchical structure model [8] and the stochastic block model [9-11]. In [12], a set of simple features are proposed as a structural model that can be analyzed to identify missing links. Hierarchical model has high accuracy in handling with the network of significant levels of the organization, such as the terrorist network and grasslands food chain network. However, since it needs to generate a lot of samples to predict the network, its computational complexity is too high to deal with the large scale networks.

Another type of link prediction method is based on the probability model. These model based methods aim at abstracting the underlying structure from the observed network, and then predicting the missing links by using the learned model. These methods first create a model containing a set of adjustable

Recebido: dia/mês/ano Aceito: dia/mês/ano

parameters, and then use optimization strategy to find the optimal parameter values, such that the resulting model can be better structures and relationships reflecting real network characteristics.

Reference [13] showed that new links can be predicted based on discovering the evolutionary model of triads in intervals between two continued snapshots of a network. The proposed algorithm in [13] is a supervised structural link prediction algorithm. If the graph is directed, there would be 64 different triads. As presented in Fig. 1, between every two nodes of a triad, there would be 4 different kinds of relations that are, one two-ways connection, two one-way connections and one no- connections. By counting these 64 different triads in two continuous snapshots of a network, a matrix called Triad Transition Matrix (TTM) can be obtained. Based on TTM matrix, the probability of a connection between two unconnected nodes can be found out.

The authors in [14] introduced a new structural supervised link prediction and analysis algorithm. The algorithm finds substructures of a graph called Vertex Collocation Profiles (VCP). If a learning phase is added to this algorithm, it can also be used for link prediction as well. The drawback of this algorithm is that it is time-consuming and unpractical for VCPs with more than 4 nodes, in large networks. Zhang and his colleagues [15], studied especial subgraphs in directed networks, and they called them microscopic organizing principles of directed networks. Their studies show that some of these subgraphs are more common in social networks. The most-favored local structure in directed networks is Bi-fan structure, which consists of 4 nodes and 4 directed links. They have proven this idea according to the homophily [16] and clustering mechanism and potential theory. Subgraphs that have only one link fewer than Bi-fan structure, that link has the highest probability to be created in the near future. This is the principal idea of the link prediction algorithm introduced in [15].

Leskovec et al. [17] developed a concept of supervised random walks. It combines the network structure with the features of nodes and edges of the network into a unified link prediction algorithm. Then they develop a method based on it. The method learns to segregate a PageRank-like random walk on the network in a supervised way, so that it is more likely to visit nodes to which new links will be create in the future. Relationship can be either positive (friendship) or negative (opposition) in social networks, a model incorporating theories of balance and status from social psychology is used to predict the signs of relationships in social networks [18]. To combine the analysis of signed networks with machine learning techniques, two categories of features are used. One is based on the degree of nodes and another is based on the principle from social psychology. Also, they investigate the network completion problem where nodes and edges in networks are both missing. They also develop KronEM, an EM approach combined with the Kronecker graphs model, to estimate the missing part of the network [19]. Moreover, Leskovec et al. collected and constructed a lot of social network datasets which are public for other researchers. These datasets have been used in many link prediction works.

Hopcroft and Tang's team [20] studies the novel problem of reciprocal relationship prediction to predict who will follow you back in directed social networks. They proposed a Triad Factor Graph (TriFG) model, which incorporated social theories (such as structural balance and homophily) over triads into the semi-supervised machine learning model. Tang's team [21] also formulated prediction problem to predict the existence and the type of links between a pair of nodes. They proposed a partially-labeled pairwise factor graph model (PLP-FGM) and two active learning strategies (Influence-Maximization Selection and Belief-Maximization Selection) to capture the inter-relationship influence [22]. They also extended the above model for the problem of inferring social ties across heterogeneous networks [23]. The model incorporates social theories into a semi-supervised learning framework, which can be used to transfer supervised information from a source network to help infer social ties in a target network. For the inventor social network where the link between inventors is the co-invention relationships. They also incorporate users's interactions into a factor graph model for recommending patent partners [24]. This method shows good prediction accuracy and

Recebido: dia/mês/ano Aceito: dia/mês/ano

efficiency, so it could be beneficial for existing recommendation models based on users's feedback.

3 Proposed Method

3.1 Feature Extraction

Node pair features are used to describe the relationship between two nodes. These features are:

A. Reciprocity and Clustering

Link reciprocity states that if there is a direct link from node n_1 to node n_2 there is a reciprocal link. Reciprocity is a very common phenomena in social networks such as friendship and coauthorship networks [25]. The clustering coefficient of a node i, C_i, is the ratio between number of triangles i it belongs to and the number of triangles that could have been formed with i as a vertex [26]. clustering coefficient between node pairs can be defined in a similar fashion [25].

B. Degree Correlation

Another node pair property considered is the node degree correlation between two nodes, assortativity. Most networks are known to exhibit either assortative or disassortative mixing. Assortativity, of a network is a Pearson correlation coefficient of the degrees at either ends of a link [25].

C. Common Neighbors

Clustering and reciprocity take shared neighbors between the node-pairs into consideration. When considering non-immediate neighbors in networks that has visibility range of greater than one hop, further information can be extracted by observing shared nodes in the neighborhood of each of the nodes. The neighborhood of node n is the set of nodes adjacent to node n and the i^{th} neighborhood of a node n is the set of nodes that are adjacent to all nodes that are in the $(i-1)^{th}$ neighborhood and nodes that do not belong to any of the previous neighborhoods [25].

3.2 Graphical Models

A Graphical model [27] is a probabilistic model for which a graph denotes the conditional independence structure between random variables.Graphical model provides a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models. In a probabilistic graphical model, each node represents a random variable, and the links express probabilistic relationships between these variables. The graph then captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors each depending only on a subset of the variables.

Graphical modeling is a powerful framework for representation and inference in multivariate probability distributions. It has proven useful in diverse areas of stochastic modeling, including coding theory [28], computer vision [29], knowledge representation [30], Bayesian statistics[31], and natural-language processing [32]. In this paper, we tackle the problem of link prediction in social network using graphical models. The next section provides details of the algorithm used for this purpose.

3.3 Markov Random Field

Markov random field (MRF) theory enables the modeling of contextual dependencies between a set of sites S. These sites might be pixels in an image or individuals in a social network. Suppose that we have a two-dimensional space, S, which has been partitioned into n nodes, labeled by the integers $A_i = \{1, 2,\ldots, n\}$ defined as state space. Each node variable can be discrete (finite or infinite) or continuous.

A lattice is a set of sites or nodes in a graph. A region with m rows and n columns can be represented as an m×n rectangular lattice, where each site corresponds to a node in the graph. An m×n lattice is written as a set of indices: $S = \{(i,j) \mid 1 \le i \le m, 1 \le j \le n\}$ or using a single index as: $S = \{i \mid 1 \le I \le m \times n\}$.

To define a Markov random field, a neighborhood structure N is needed, which defines the range of interaction from one node to

Recebido: dia/mês/ano Aceito: dia/mês/ano

another. A neighborhood system $N = \{ N_i, \forall i \in S \}$ is a collection of subsets of S for which

1) $i \notin N_i$ (a site is not part of its neighborhood) and

2) $j \in N_i \Leftrightarrow i \in N_j$ (i is in the neighborhood of j if and only if j is in the neighborhood of i).

In general, $\forall s \in S : s = (i , j)$, an n^{th}-order homogeneous neighborhood system could be defined as $N^n = \{N_{(i,j)}: (i,j) \in S\}$ and $N^n_{(i,j)} = \{(k,l) \in S : (k-i)^2 + (l-j)^2 \leq n\}$.

A clique C is a subset of S for which every pair of sites is neighbors. Single node is also considering cliques. The set of all cliques on a lattice is called C. A random field, with respect to a neighborhood structure, is a Markov random field if the joint probability density on the set of all possible intensity values x satisfies the following properties:

1) $p(X) > 0$ for all X

2) $p(\text{all nodes in the lattice except } x_i) = p (X_i | \text{ neighbors of } x_i)$[33].

The definition of Markov random fields stated that the probability measure must fulfill the local Markov property. This is not a restriction for the equivalence between local and global Markov property since it is considered the following theorem. Thus the joint probability is given by $P(x) = \frac{1}{Z} \exp\{-U(x)\}$, where Z is the normalized constant or partition function and U(x) is the energy function with form $U(x) = \sum V_c(X)$ with the summation that is over the local clique potentials set $V_c(x)$ over all possible cliques C[29].

4 Experimental Results

4.1 Dataset

The purpose of this paper is to predict friendship relationship with high probability. This prediction helps social network websites a lot in finding out the existence of a relation between two individuals. To do so, we used the data of Kaggle competition site which have been collected from Flickr social site. Flickr is a huge social network having 36 millions of users and 35 billions of photos. This site is full of friendship data, including people's comments, group memberships, friend suggestions, clicking on favorite's photo, and restricting the visit to some of the friends and families. Data consist of two test and train files which we analyzed separately in the following:

The first file, ''Social_Train.zip'' contains of 7,237,983 records with two columns of first person and second person. These columns are filled numerically which denotes person unique number that are assigned to a person within the whole data. There are 1,133,574 different individuals in the data. Each column shows that the first person is friend with the second person. The second file is ''Social_Test.zip'' which includes8,960 records and three columns, like the first file, it has two columns of first and second person, and the third column is the prediction column which represents whether or not the first person and the second person are friends. These columns are filled with 0 and 1, value 1 in the case of friendship existence, and value 0 otherwise. These data have been collected from December 2010 to January 2011.

4.2 Evaluation Method And Criteria

In this paper, we exploited ROC curve to compute the validity of predicted values. ROC is a strong simulation tool which is used in medical decision making, psychology, communications and whenever need for threshold values is concerned (Zouet al. 2007).Entries will be evaluated using the area under the receiver operator curve (AUC).Today, it is a commonly used evaluation method for binary chooses problems, which involve classifying an instance as either positive or negative. To understand the calculation of AUC, a few basic concepts

must be introduced. For a binary choice prediction, there are four possible outcomes, Table 1.The true positive rate, or recall, is calculated as the number of true positives divided by the total number of positives. The false positive rate is calculated as the number of false positives divided by the total number of negatives.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Table 1: AUC calculation guide

True positive (TP)	A positive instance that is correctly classified as positive
False positive(FP)	A negative instance that is incorrectly classified as positive
True negative (TN)	A negative instance that is correctly classified as negative
False negative(FN)	A positive instance that is incorrectly classified as negative

We also use accuracy, precision and recall as evaluation metrics for evaluation of the proposed method according to the following relationships:

accuracy = (TP + TN) / (TP + TN + FP + FN)

precision = TP / (TP + FP)

recall= TP / (TP + FN)

4.3 Result And Discussion

Rehelp method [34] and Decision Tree algorithm [35], which has been reported to significantly improve on previous approaches, are used to compare with the proposed method. To compare the proposed method to each method, we provide experiments to generate values for the evaluation metrics from two methods and statistically tested significance of differences using a paired t –test. The learning and testing algorithms were implemented in Matlab. We used 10 fold cross-validation (CV) repeated 10 times to generate results for evaluation. The comparative results are shown in Fig. 1 with details in Table 2.

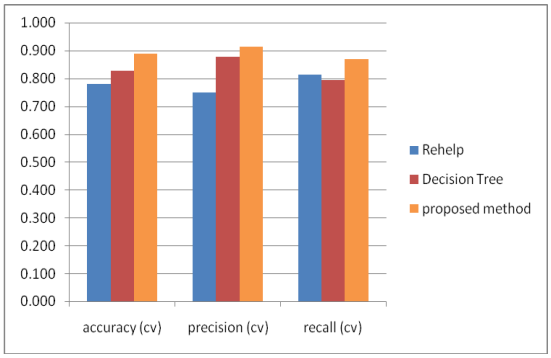

Figure 1: Comparative Results. CV indicates10-fold cross validation.

Table 2 represent the accuracy, precision and recall for each method. With regard to the obtained results, we gained a complete graph which compares predicated and existing data for each method based on ROC criterion and shows the ROC curve of the proposed method remains above that of two methods [34], [35] shown in figure 2.

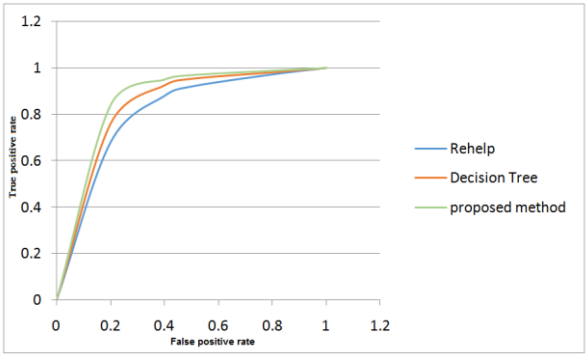

Figure 2: ROC for comparison of the proposed method with Rehelp method [34] and Decision Tree algorithm [35].

Table 2: Comparative Results between proposed mehod with Rehalp [34] and Decision Tree [35]

		Rehelp [34]	DecisionTree [35]	Proposed MRF
10-fold CV	Accuracy	0.781	0.826	**0.889**
repeated 10 times	Precision	0.749	0.877	**0.915**
	Recall	0.813	0.795	**0.870**

Recebido: dia/mês/ano Aceito: dia/mês/ano

5. Conclusion and Future Work

The objective of this paper was to present a method for link prediction in social network using Markov random field. For this purpose, the Flickr website data available on Kaggle website were used. The data consist of train and test files. By exploiting the first file as train data, the relationships between individuals were determined in the second file. The experiments on Flickr datasets show that our method has better performance of link prediction than other methods in the typical networks like social networks, and the MRF algorithm can improve the accuracy of link prediction.

In the future work, on the one hand, we will examine and test more datasets from other domains,. In the future, we would like to extend our experimental datasets to larger ones from other sources to test the effectiveness and robustness of our proposed method.

References

[1] Nettleton, D.F. (2013). Data mining of social networks represented as graphs. Computer Science Review, vol. 7, p. 1-34.

[2] Sherkat, E., Rahgozar, M., Asadpour, M. (2015). Structural link prediction based on ant

colony approach in social networks. Physica A: Statistical Mechanics and its Applications, vol. 419, p. 80-94.

[3] Sharma, U., Sharma, D., Khatri, S. K. (2015). Elimination based algorithm for link prediction on social networks. International Journal of System Assurance Engineering and Management, vol. 6, no. 1, p. 78-82.

[4] He, Y., N. K. Liu, J., Hu, Y., Wang, X. (2015). OWA operator based link prediction ensemble for social network. Expert Systems with Applications, vol. 42, no. 1, p. 21-50.

[5] Schall, D. (2014). Link prediction in directed social networks. Social Network Analysis and Mining.

[6] Chen, B., Chen, L. (2014). A link prediction algorithmbased on ant colony optimization. Applied Intelligence, vol. 41, no. 3, p. 694-708.

[7] Aiello, L. M., Barrat, A., Schifanella, R. (2012). Friendship prediction and homophily in social media. Journal ACM Transactions on the Web (TWEB), vol. 6, no. 9.

[8]Clauset, A., Moore, C., Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. Nature, vol. 453, p. 98-101.

[9] White, H. C., Boorman, S. A., Breiger, R. L. (1976). Social structure from multiple networks. I. block models of roles and positions. *American Journal of Sociology, vol. 81, no. 4, p.* 730-780.

[10] Doreian, P., Batagelj, V., Ferligoj, A. (2004). Generalized blockmodeling with Pajek. Cambridge University Press, vol. 1, no. 2, p. 455-467.

[11] Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, X. P. (2008). Mixedmembership stochastic block models. Journal of Machine Learning Research.

[12] Fire, M., Tenenboim, L., Lesser, O. (2011). Link prediction in social networks using computationally efficient topological features. IEEE third international conference on privacy, security, risk and trust (passat), p. 73–80.

[13] Juszczyszyn, K., Musial, K., Budka, M. (2011). Link prediction based on subgraph evolution in dynamic social networks. in: Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on SocialComputing (SocialCom), p. 27–34.

[14] Lichtenwalter, R. N., Chawla, N. V. (2012). Vertex collocation profiles: subgraph counting for link analysis and prediction. the International World Wide Web Conference Committee (IW3C2), p. 1019–1028.

[15] Zhang, Q. M., Lü, L., Wang, W. Q., Zhou, T., Xiao, Y. (2013). Potential theory for directed networks. PLoS One.

[16] McPherson, M., Smith-Lovin, L., Cook, J. M. (2001). Birds of a feather: homophily in social networks. Annual Review of Sociology, vol. 27, p. 415-444.

Recebido: dia/mês/ano Aceito: dia/mês/ano

[17] Backstrom, L., Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11), p. 635–644.

[18] Leskovec, J., Huttenlocher, D., Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10), p. 641–650.

[19] Kim, M., Leskovec, J. (2011). The network completion problem: inferring missing nodes and edges in networks. In: Proceedings of the 11th SIAM International Conference on Data Mining (SDM'11), Mesa, p. 47–58.

[20] Hopcroft, J., Lou, T., Tang, J. (2011). Who will follow you back? reciprocal relationship prediction. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM11), Glasgow, p. 1137–1146.

[21] Tang, W., Zhuang, H., Tang, J. (2011). Learning to infer social ties in large networks. In: Proceedings of Machine Learning and Knowledge Discovery in Databases European Conference (ECML/PKDD), Athens, vol. 6913, p. 381–397.

[22] Zhuang, H., Tang, J., Tang, W., et al. (2012). Actively learning to infer social ties. Data Mining and Knowledge Discovery, vol. 25, p. 270–297.

[23] Tang, J., Lou, T., Kleinberg, J. (2012). Inferring social ties across heterogeneous networks. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12), Seattle, p. 743–752.

[24] Wu, S., Sun, J., Tang, J. (2013). Patent partner recommendation in enterprise social networks. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13), Rome, p. 43–52.

[25] Feyessa, T., Bikdash, M., Lebby, G. (2011). Node-Pair Feature Extraction for Link Prediction. Department of Electrical and Computer Engineering North Carolina A&T State University Greensboro, North Carolina 2741, 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing.

[26] Fagiolo, G. (2007). Clustering in complex directed networks. Physical Review E.

[27] Koller, D., Friedman, N. (2009). Probabilistic graphical models. MITpress.

[28] McEliece, R. J., MacKay, D. J. C., Cheng, J. F. (1998). Turbo decoding as an instance of Pearl's belief propagation algorithm. IEEE Journal on Selected Areas in Communications, vol. 16, no. 2, p. 140–152.

[29] Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, p. 721–741.

[30] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann Series in Representation and Reasoning).

[31] Gelfand, A. E., Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, vol. 85, p. 398–409,.

[32] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, vol. 3, p. 993-1022.

[33] Zimeras, S. (2012). Markov Random Field and Social Networks. Virtual Communities, Social Networks and Collaboration, Annals of Information Systems, vol. 15, p. 207-219.

[34] Cai, X., Bain, M., Krzywicki, A., et al. (2012). Reciprocal and Heterogeneous Link Prediction in Social Networks. Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, vol. 7302, p. 193-204.

[35] Al Hasan, M., Chaoji, V., Salem, S., Zaki, M. (2006). Link Prediction using Supervised Learning. In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security.

Recebido: dia/mês/ano Aceito: dia/mês/ano