A Review of Link Prediction in Social Networks

Tingli Wang
School of Information Technology
Jiangxi University of Finance and Economics
Nanchang, China
wangtingli.jiayou@163.com

Guoqiong Liao
School of Information Technology
Jiangxi University of Finance and Economics
Nanchang, China
liaoguoqiong@163.com

***Abstract*—In recent years, with the emerging of social media, such as Facebook, Twitter, Sina Microblog, more and more researchers pay their attentions to social networks. Link prediction is one of the most interesting issues among the social network analysis, which exploits existing networks information, like the characteristics of the nodes and edges, to predict potential relationships to be formed in the future. This paper summarizes the link prediction methods in social networks, including tradition link prediction methods, link prediction methods in heterogeneous networks, and temporal link prediction methods.**

Keywords-link prediction; socail networks; heterogeneous networks

I. INTRODUCTION

In recent years, with the emerging of social media, such as Facebook, Twitter, Sina Microblog, more and more researchers pay their attentions to social networks. A social network looks like a graph structure consisting of nodes and edges, where the edges represent the links, i.e. the interactions, between the related nodes. The links play important roles in social network analysis, since many potential social phenomena in the networks can be revealed by analyzing the relationships among the links. Essentially, the social networks are highly dynamic and changeable. since it is nature that the old objects may quit and the newcomers will join.

Link prediction is one of the most interesting issues among the social network analysis, which exploits existing networks information, like the characteristics of the nodes and edges, to predict the potential links to be formed in the future[1~2]. Link prediction can be used for recommending commodities, discovering the missing links, and identifying the false links, etc[3~4]. Therefore link prediction is attracting more and more attentions of the experts in different fields.

The exiting link prediction methods are mostly designed for static and homogeneous networks, that is, the types of nodes or edges are single. But the these prediction models cannot be used for dynamic and heterogeneous networks, since they haven't considered the dynamic feature of the networks, Generally, the real social networks are complicated, and a network may has different types of nodes or links, which may have different contents or attributes. For example, in DBLP network, a node may be an author, a conference or a paper, and a link relationship may be a co-author or a paper–writing relationship. Han et al. [5] defined these complex networks with multiple types of nodes or links as heterogeneous information networks.

The paper intends to summarize the link prediction methods in social networks. Section II gives problem definition of link prediction. In Section III, tradition link prediction methods are discussed. Section IV introduces the link prediction methods in heterogeneous networks. In Section V, temporal link prediction methods are analyzed. Section VI is the summary of the paper.

II. PROBLEM FORMULATION

Given a social network $G=(V, E)$, where $V=\cup V_i$ (i=1,2,…, n) is the union of different node sets, $E=\cup E_j$ (j=1, 2, …, m) is the union of different link sets. If $i>1$ or $j>1$, the network is a heterogeneous network, otherwise, is a homogeneous network.

The general task of link prediction in the social networks is to predict whether it will form a new link with a specific type between a given node pair $x, y \in V$ in the future.

Unlike static link prediction, temporal link prediction should consider dynamic evolution of the networks. Given the nodes and the links in time T, the task of temporal link prediction is to predict the links at time T+1.

Generally, both unsupervised learning methods and supervised learning methods are used to do link prediction. When unsupervised learning methods are used, the problem becomes how to assign scores to the pairs of nodes for the target link type. If the score is more than a given threshold, a link is likely to be formed between the pair. While in the case of supervised learning, link prediction is typically converted into a classification problem by extracting features from training data.

III. TRADITIONAL LINK PREDICTION

A. Methods Based on Similarity Measure

Due to the importance of link prediction in practical applications, a lot of link prediction methods have been proposed. The most widely used methods are based on similarity (proximity) measures, which assume the higher the measure score is, the more likely a link will be formed. As we know, there are many ways to define the similarity based on the information of node attributes and network structures. But due to privacy protection, reliable node attributes is hard to be obtained, so the methods based on the similarity of network structures are widely used. The critical of structure similarity is whether it can well seize the characteristics of the target networks.

978-1-4799-6543-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMeCG.2014.38

Liben-Nowell et al. [7] firstly gave the definition of link prediction. They summarized link prediction methods by computing similarity based on node neighborhoods or the ensemble of all paths between the pairs of nodes.

There are many methods based on the similarity of node neighborhoods, such as Common Neighbor (CN) [8], Jaccard's Coefficient (JC) [9], Adamic/Adar (AA) [10], Preferential Attachment (PA) [11], etc. For each pair of nodes, the more the common neighbor number is, the higher the similarity score will be. These algorithms have been applied to a lot of networks, and experiments show that they can achieve better performance.The simplest but efficient approach is based on Common Neighbor, which mathematical expression is:

$$CN(x,y) = \left|\Gamma(x) \cap \Gamma(y)\right| \tag{1}$$

where $\Gamma(x)$ and $\Gamma(y)$ denote the neighbor set of node x and node y respectively.

Another important algorithm related to the network topology is Katz [12]. It directly sums over all the paths that exist between a pair of vertices. But, to penalize the contribution of longer paths in the similarity computation it exponentially damps the contribution of a path.

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l \cdot \left|paths_{x,y}^l\right| \tag{2}$$

where $paths_{x,y}^l$ is the set of all paths with length l connecting x and y, and β is the damping factor controlling the path weights. M is the matrix of the network, the similarity matrix can be written as:

$$Katz(x,y) = (I - \beta M)^{-1} - I \tag{3}$$

However, the computational complexity of Katz is higher than others, since it has to consider the global information.

In addition, Lichtenwalter et al. [13] proposed a new unsupervised learning method PropFlow, which is based on paths. PropFlow is similar to Rooted PageRank, but it is more localized measure of propagation, and is insensitive to topological noise far from the source nodes.

B. Methods Based on Classification Model

The link prediction problem can also be regarded as a classification problem. Comparing with the similarity measures, the classification model is more flexible, since the features extracted by similarity-based methods can be directly used to train the classification models for link prediction.

Hasan et al. [14] pointed out that extracting feature set depends on the actual application. They compared the results of diverse classification algorithms, including SVM, decision trees, K-nearest neighbor, etc. The result shows that SVM is better than others. Benchettara et al. [15] apply supervised machine learning to predict the links in two-mode social networks. However, these studies don't consider the dynamic evolution and diversity of the networks.

Although the classification model can improve performance, how to choose the appropriate features is still a problem. Moreover, if the networks are sparse, the positive and negative classes for training are extremely imbalance, it will seriously affect the accuracy of link prediction.

C. Methods Based on Probabilistic Model

Wang et al. [16] proposed a local probabilistic model for link prediction that uses an undirected graphical model, Markov Random Fields, to model the local neighborhood containing two nodes. In order to reduce the cost for large scale networks, they build a local model rather than a global model. There are two stages in the approach, i.e., firstly to determine the nodes that will be included in the local model, then to use frequent non-derivable itemsets to determine the structure of the graphical model as well as learning the parameters of the model.

In addition, Kashima et al. [17] present a novel parameterized probabilistic model of network evolution and derive an efficient incremental learning algorithm for such models, which is used to predict links among the nodes.

The probabilistic model can depict the relationships of network structures, fully utilize the network information, and has higher prediction accuracy than others. However, the model is limited by the computational complexity, therefore it is not suitable for large-scale networks.

IV. HETEROGENEOUS LINK PREDICTION

Buy now, there were abundant researches focusing on inferring particular types of links in homogeneous social networks, while few publications systematically study the problem to infer the links over the heterogeneous networks.

A. Usual Link Prediction

From the view of existing research, there are two typical ways of handling link prediction for heterogeneous networks: (1) treat all types of links equally; (2) study each type of link separately and ignore the correlation between link types [4].

Davis et al. [6] extend triad census to the heterogeneous networks, and count the occurrence of each triad census as weights of link types. They proposed an unsupervised method MRLP, which is a weighted extension of Adamic/Adar. It is proved that MRLP is superior to the traditional link prediction methods. For the unsupervised learning method is not flexible, Davis also developed a multi-relational supervised learning method MR-HPLP. In the work, link types are not treated equally. Triad census will have high computational complexity in large-scale networks.

Yang et al. [23] designed an efficient link prediction model in the heterogeneous networks. Differing from MR-HPLP, Yang quantifies the correlation between different types of links from the view of influence propagation, and proposed a new unsupervised method MRIP, to reduce the computational complexity. The design of MRIP is based on two considerations [23]: (1) For any given link type i, the influence propagates not only through the links of type i but also propagates through other types of links. (2) The

probabilities that propagate through other link type j depend on the correlation between link type i and link type j.

The influence score as follows:

$$flow(x,y,i) = score(x)\cdot\beta\cdot\frac{weight(x,y,i)}{\deg ree(x,i)} + score(x)\cdot\beta\cdot\sum_{j\neq i}^{K}\left(\sigma(i,j)\cdot\frac{weight(x,y,i)}{\deg ree(x,i)}\right)/\left(\left|E(x,y)\right|-1\right) \quad (4)$$

where x and y are nodes, β is the Katz factor, $score(x)$ is the probability of a link between the source node and node x, and $|E(x,y)|-1$ is the number of link types between node x and y except type i.

In addition, Aggarwal et al. [24] gave a framework for dynamic link prediction in the heterogeneous networks.

In real social networks, out of concern over privacy issues, a lot of useful information cannot be obtained. Therefore, in the case of lack of detail information, Kuo et al. [25] devised a novel unsupervised framework to predict the opinion holder in a heterogeneous social network without any labeled data. By constructing a three-layer factor graph model, employing effective learning and inference algorithm, the approach wins baseline approaches.

B. Anchor Link Prediction

With the emerging of social networks, people often have multiple accounts in the social networks among different websites. In other words, the same person may involve in multiple networks. Kong et al. [26] devoted themselves to mining the correspondence between these accounts across multiple social networks, which are defined as the anchor links. On the premise of known part of the anchor link, they extract heterogeneous features from multiple heterogeneous networks, and derived a solution Multi-Network Anchoring (MNA) to infer anchor links.

C. Link Prediction for New Users

Unlike the general users, new users are often lack of information and have different characteristics with the old users, which make it difficult to predict the links of the new users. Zhang et al. [27] propose a supervised learning method called SCAN-PS (Predicting Links across Aligned Networks with Personalized Sampling), which exploit information from the source networks for link prediction in the target networks. The method can solve the cold start problem, and experiment shows that it outperforms other link prediction methods for new users.

V. Temporal Link Prediction

In the real world, most of social networks are dynamic and changeable, which bring more challenges for link prediction.

A. Methods Based on Matrix and Tensor Decomposition

Daniel et al. [18] present two temporal link prediction methods, matrix-based and tensor-based, for bipartite graphs. The co-authorship network is a typical bipartite network, where the types of nodes are authors and conferences. They also show how the Katz method is extended to bipartite graphs, and use the truncated SVD to devise a scalable method for calculating a "truncated" Katz score. They also illustrate the usefulness of the natural three-dimensional structure of temporal link data by using CANDECOMP/PARAFAC (CP) tensor decomposition of the data. The drawback of the tensor-based approach is that there is a higher computational cost incurred.

B. Methods Based on Time Series Model

Huang et al. [19] briefly discuss commonly used link prediction algorithms under a static graph representation, and present their method based on time series model ARIMA. Despite the dynamic evolution of network will be considered, however, this model does not consider the temporal correlations among the links and the large number of links makes such models intractable. In addition, they propose a hybrid prediction algorithm, i.e., a combination of algorithms under the static graph representation and time series models. Experiments show the method has better performance than the six commonly used static graph link prediction algorithms.

Paulo et al. [20] tried to deal with the limitation of traditional approaches by exploring topological metrics in the networks over time. They addressed the evolution of the networks as a time series problem. Firstly, By using the basic similarity measure algorithms, such as PA, CN, etc., similarity metrics over time of the network are calculated. Then, they used a set of well known statistical forecasting models, such as Moving Average, Random Walk, and so on, to estimate future values. The final value will be used in the unsupervised learning and supervised learning models.

The application of time series model is limited, since the method can only be applied to the situation where there are abundant time series link occurrence.

C. Other Temporal Methods

Due to the expression of traditional methods is not sufficient to describe the dynamic features, Paulo et al. [21] proposed an unsupervised predictive measure. They defined a temporal event as a specific activity between two nodes from a frame to its subsequent. The proposed measure combines the rewards associated to primary events, which are the temporal events strictly related to the pair of nodes under analysis, and the rewards associated to secondary events, which are the temporal events observed in the nodes' neighborhood. The proximity score associated with a given pair of nodes (x, y) defined as:

$$score(x,y) = \sum_{k=2}^{n}\beta(k)\left[P(x,y,k)+\alpha S(x,y,k)\right] \quad (5)$$

Where $P(x, y, k)$ is the reward of the event (conservative, innovative or regressive) for the pair of nodes (x, y) observed in the transition from frame k-1 to frame k, and $S(x, y, k)$

indicates the aggregated reward of secondary events associated to the pair (x, y).

Catherine et al. [22] put forward a Covariance Matrix Adaptation Evolution Strategy (CMA-ES), which is, a linear combination of sixteen neighborhoods and node similarity indices. The algorithm is flexible and allows for any similarity index to be substituted into or added to the evolutionary algorithm. But simply a linear combination of these indicators will not meet the practical significance. These methods take temporal information into account, but without considering the complex situation of social networks.

VI. SUMMARY

Social network analysis has attracted more and more attentions of researchers in various fields. The paper introduces the concept of link prediction and summarizes the latest researches and methods of link prediction. Even though previous studies have made certain achievements in link prediction, especially the similarity measures, but there are still a lot of problems to be solved as follows.

1) The sizes of social networks are usually very huge, which makes it is extremely hard to study and analysis the social network.

2) The information of some social networks is sparse, which makes supervised link prediction methods suffer from the problem of class imbalance.

3) There are still few researches considering the dynamic and heterogeneous characteristics of the networks, it is still necessary to design more appropriate methods for dynamic link prediction in the heterogeneous networks.

ACKNOWLEDGMENT

This work is supported partly by NSF of China (No. 61262009), NSF of Jiangxi, China (20122BAB201032), Science Foundation of Jiangxi Provincial Department of Education, China (No. GJJ10694, GJJ12259), Building Project of Superior Innovation Team of Science and Technology of Jiangxi, China (No. 20113BCB24008).

REFERENCES

[1] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

[2] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[3] Huang Z, Li X, Chen H. Link prediction approach to collaborative filtering[C]//Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2005: 141-142.

[4] Dong Y, Tang J, Wu S, et al. Link prediction and recommendation across heterogeneous social networks[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012: 181-190.

[5] Han J. Mining heterogeneous information networks by exploring the power of links[C]//Discovery Science. Springer Berlin Heidelberg, 2009: 13-30.

[6] Davis D, Lichtenwalter R, Chawla N V. Multi-relational link prediction in heterogeneous information networks[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011: 281-288.

[7] Liben‐Nowell D, Kleinberg J. The link‐prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031.

[8] Newman M E J. Clustering and preferential attachment in growing networks[J]. Physical Review E, 2001, 64(2): 025102.

[9] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura[M]. Impr. Corbaz, 1901.

[10] Adamic L A, ADAR E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230.

[11] Xie Y B, Zhou T, Wang B H. Scale-free networks without growth[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(7): 1683-1688.

[12] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.

[13] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 243-252.

[14] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. 2006.

[15] Benchettara N, Kanawati R, Rouveirol C. Supervised machine learning applied to link prediction in bipartite social networks[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on. IEEE, 2010: 326-330.

[16] Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction[C]//Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007: 322-331.

[17] Kashima H, Abe N. A parameterized probabilistic model of network evolution for supervised link prediction[C]//Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006: 340-349.

[18] Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2011, 5(2): 10.

[19] Huang Z, Li X, Chen H. Link prediction approach to collaborative filtering[C]//Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2005: 141-142.

[20] da Silva Soares P R, Bastos Cavalcante Prudencio R. Time Series Based Link Prediction[C]//Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, 2012: 1-7.

[21] Soares P R S, Prudêncio R B C. Proximity measures for link prediction based on temporal events[J]. Expert Systems with Applications, 2013, 40(16): 6652-6660.

[22] Bliss C A, Frank M R, Danforth C M, et al. An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks[J]. arXiv preprint arXiv:1304.6257, 2013.

[23] Yang Y, Chawla N, Sun Y, et al. Predicting Links in Multi-relational and Heterogeneous Networks[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012: 755-764.

[24] Aggarwal C C, Xie Y, Yu P S. A framework for dynamic link prediction in heterogeneous networks[J]. Statistical Analysis and Data Mining, 2013.

[25] Kuo T T, Yan R, Huang Y Y, et al. Unsupervised link prediction using aggregative statistics on heterogeneous social networks[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 775-783.

[26] Kong X, Zhang J, Yu P S. Inferring anchor links across multiple heterogeneous social networks[C]//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013: 179-188.

[27] Zhang J, Kong X, Yu P S. Predicting Social Links for New Users across Aligned Heterogeneous Social Networks[J]. arXiv preprint arXiv:1310.3492, 2013.