Link prediction in social network based on local information and attributes of nodes

Yingying Liang\textsuperscript{1, 2}, Lan Huang\textsuperscript{1, 2, 3} and Zhe Wang\textsuperscript{1, 2, *}

\textsuperscript{1}College of Computer Science and Technology, Jilin University, 130012, Changchun, China,
\textsuperscript{2}Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, 130012, Changchun, China
\textsuperscript{3}Zhuhai College of Jilin University, 519041, Zhuhai, China
\ *Corresponding Author: wz2000@jlu.edu.cn

Abstract. Link prediction is essential to both research areas and practical applications. In order to make full use of information of the network, we proposed a new method to predict links in the social network. Firstly, we extracted topological information and attributes of nodes in the social network. Secondly, we integrated them into feature vectors. Finally, we used XGB classifier to predict links using feature vectors. Through expanding information source, experiments on a co-authorship network suggest that our method can improve the accuracy of link prediction significantly.

1. Introduction

Link prediction plays an important role in the field of data mining [1]. The traditional methods are based on the topological information, which have been widely used in various fields [1], like social recommendation, information retrieval and other fields. It is also very helpful to solve the problem of information overload [2][3]. The traditional methods can be divided into two categories, the first category is based on the local information like CN [4], and the other one is based on the global information like Katz [5].

Centrality is used to describe importance of nodes in the network [6][7]. If the centrality of a node is higher than others, it is easier for it to attract attention and construct new links with other nodes in the network, so researchers improve accuracy of link prediction methods with the help of centrality[8]. Some researchers construct feature vectors using characteristics including centrality of nodes to predict links[9][10], and some researchers use machine learning methods to improve performance of link prediction methods[11].

It has become more convenient for people to communicate with others using social media websites which produced a huge amount of information. In order to improve the accuracy of link prediction, researchers begin to pay attention to the relevance between content and structure of the social network, and then design new methods to predict links effectively [12][13]. Many researchers used some characteristics of social networks to improve accuracy of link prediction, and got good results[14][15]. According to [12], people belong to the same social network tends to form a local community, so they defined a new index: CAR which can improve the effect of classical CN index significantly. CAR is also better than AA index [14] and RA index [15].

The relationship in social networks is dynamic with the change of time and network context [9]. In a co-authorship network, if two authors have no co-authorship at the moment, how can we predict whether they will construct relationship in the future? In this paper, we focus on topology and context
of network and consider attributes of nodes, then design feature vectors to describe characteristics of nodes. In addition, this kind of framework is suitable for most social networks since the calculation of features is very easy. Our method is based on classification using XGB classifier, and experiments on a part of DBLP (Digital Bibliography Library Project) show that our method can predict links effectively.

2. Related Definitions and Concepts

In order to formalize characteristics of the social network, researchers usually use graph theory to model it. For a social network, each individual is considered as a point, and the relationship between them is considered as an edge, here we will give some basic definitions.

Definition 1. Social network. \( G = (V, E) \) represents a social network, where \( V \) represents the set of users, \( E \) represents the set of edges, and \( e_{i,j} \in E \) represents an edge between nodes \( v_i \) and \( v_j \).

Definition 2. \( CN \) index. For any network \( G \), if there is no direct link between \( v_i \) and \( v_j \) at the moment, then \( CN(v_i, v_j) \) index (Common Neighbors) is defined as in equation (1), where \( \Gamma(v_i) \) represents the set of neighbors of \( v_i \), and \( \Gamma(v_j) \) represents the set of neighbors of \( v_j \).

\[
CN(v_i, v_j) = \Gamma(v_i) \cap \Gamma(v_j)
\]

(1)

Definition 3. \( CAR \) index. If nodes \( v_i \) and \( v_j \) have no direct edges at the moment, then \( CAR \) index [12] is defined as in equation (2), where \( LCL(v_i, v_j) \) represents the number of edges between nodes which belong to \( CN(v_i, v_j) \). \( CAR \) index can improve the accuracy of \( DC \) index effectively without increasing too much computation.

\[
CAR(v_i, v_j) = CN(v_i, v_j) \times LCL(v_i, v_j)
\]

(2)

Definition 4. \( DC \) index. \( DC \) index (Degree Centrality) represents centrality of a node using the number of neighbors of it. Taking an example of \( v_i \), \( DC(v_i) \) is defined as in equation (3), where \( k(v_i) \) represents the number of neighbors of \( v_i \), and \( N \) is the total number of nodes in the network, so \( N - 1 \) represents the maximum number of neighbors of a node in the network.

\[
DC(v_i) = \frac{k(v_i)}{N - 1}
\]

(3)

We choose XGB classifier [16] to predict links. Babajide [17] proposed an algorithm to predict molecular activities based on XGB and experiments show that it is more accurate than some popular algorithms at present, like support vector machine, stochastic forest, naive Bayesian and so on.

3. Methods and Experiments Setup

Common neighbors of two users often have an impact on the construction of new links between strangers, so we propose the concept of relevant force to describe influence of \( CN \) nodes at first and design the modified \( CN \) index by combining relevant force. Then we design some other features to describe attributes of nodes. Finally, we construct feature vectors using those features to predict links.

3.1. Our Method

3.1.1. Relevant force (\( RF \)). \( RF \) calculates \( CN \) nodes’ influence according to specific context, if there is a link between \( v_j \) and \( v_z \), then \( RF \) from \( v_z \) to \( v_j \) is defined as in equation (4):
\[ RF(v_i, v_z) = \frac{1 + CN(v_i, v_z)}{D(v_i) + \sum_{m=1}^{D(v_i)} CN(v_i, v_m)} (v_z \in CN(v_i, v_j), v_m \in Nei(v_j)) \]  

(4)

where \( \sum_{m=1}^{D(v_i)} CN(v_i, v_m) \) represents the number of common neighbors between \( v_i \) and its neighbors, and \( D(v_i) \) represents degree of \( v_i \).

3.1.2. **MCN index.** We assume that there is no direct link between \( v_i \) and \( v_j \), and \( v_z \) is one of nodes of \( CN(v_i, v_j) \), then \( RF \) index is defined as in equation (5):

\[ MCN(v_i, v_j) = \sum_{v_z \in f'(v_i) \cap f'(v_j)} ((RF(v_i, v_z) + RF(v_j, v_z)) \times DC(v_z)) \]  

(5)

For any node belongs to \( CN(v_i, v_j) \), there are two factors affect the influence from it to \( v_i \) and \( v_j \): its importance in local network and the trust degree from \( v_i \) and \( v_j \); to it. We measure them through \( DC \) and \( RF \) respectively. The higher value of \( MCN \) means that there is greater possibility to construct new links.

3.2. **Experiments Setup**

DBLP is a collection of English literatures in the field of computer science which lists information of authors and papers. Link prediction on DBLP refers to predicting co-author relationship between authors in the future based on information within specific period. The author of the paper corresponds to a node, and the co-author relationship corresponds to the edge. We pay attention to authors’ research fields, and integrate co-authorship information into \( MCN \). Here are some definitions based on DBLP.

3.2.1. **Similarity of research area.** In the field of computer science, different international academic conferences focus on different research areas. We select 15 international conferences in DBLP and divide them into four areas. The details as follows: 1) Data mining: KDD, ICDM, PAKDD; 2) Database: SIGMOD, VLDB, ICDE, PODS; 3) Information retrieval: SIGIR, ECIR, WWW, CIKM; 4) Machine learning: ICML, NIPS, AAAI, CVPR.

We assume that if the author published a paper at an international conference which belongs to a specific research area, there will be a link between the author and the research area. Then we regard the number of links between authors and research areas as a feature which expresses author’s interests. We use cosine similarity of feature vectors to represent the similarity between authors. We regard \( A = (A_1, A_2, ..., A_n) \) and \( B = (B_1, B_2, ..., B_n) \) as the authors’ feature vectors, then the \( CosSim \) (Cosine Similarity) is defined as in equation(6).

\[ CosSim(A, B) = \frac{\sum_{i=1}^{n} (A_i \times B_i)}{\left(\sum_{i=1}^{n} A_i^2\right)^{1/2} \times \left(\sum_{i=1}^{n} B_i^2\right)^{1/2}} \]  

(6)

We define the vector \( Confer \) to represent author’s interests, so similarity of research areas is defined as \( ConfSim \) which is calculated as in equation (7):

\[ ConfSim(v_i, v_j) = CosSim(Confer(v_i), Confer(v_j)) \]  

(7)

3
3.2.2. Modified MCN index. In DBLP, authors who have not direct links may construct a link if they can communicate with each other through a common neighbor. Researchers often construct new links with each other due to similar research topics. We define CoPub to describe the times of co-author between authors, and then define modified MCN index: MMCN using CoPub and RF. MMCN is defined as in equation (8):

\[
MMCN(v_i, v_j) = \sum_{r_i \in CN(v_i, v_j)} DC(v_i) \times \sum_{k=i,j}(RF(v_i, v_j) \times CoPub(v_i, v_j))
\] (8)

Unlike the MCN index, CoPub and ConfSim are the indexes which are designed according to the characteristics of DBLP. According to the specific conditions, we can extract different kinds of domain knowledge from different fields, and integrate them to the feature vectors, all of which will be helpful to improve the accuracy of link prediction methods.

4. Results and Discussion

After obtaining the information such as network topology and attributes of nodes, we can calculate the CAR and other indexes, and then construct the feature vector to predict links. Let the feature vector as Features = \{f_1, f_2, ..., f_n\} where f_i represents various specific indexes 1CAR, DC, MCN and so on. We use the indexes presented in this paper to construct a model of feature vector and test it on DBLP.

4.1. Data Set

We select a part of DBLP and define it as DBLP_fourarea which contains 1500 authors and 6303 co-author relationships. It includes information of authors and papers published in 15 international conferences mentioned in 3.2.1 during the period from 2004 to 2010. Then we select information of DBLP_fourarea from 2004 to 2008 as the train set, and regard the information from 2009 to 2010 as the test set. Our aim is to predict co-author relationships which are included in train set but not exist in test set.

4.2. Evaluation Standards

In this paper, we choose F-value and AUC to evaluate the performance of our method. F-value combines precision and recall, which is the harmonic mean of precision and recall, so it is more reliable. The traditional F-value is defined as in equation (9):

\[
F-value = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\] (9)

The area under the curve (AUC) describes the ability of the classifier to classify the patterns submitted correctly. It is defined as in equation (10):

\[
AUC = \frac{1}{2} \sum_{i=1}^{n-1} (x_i + y_i + y_{i+1})
\] (10)

where \(x_i\) and \(y_i\) represent the X-axis and Y-axis of the ROC curve respectively.

4.3. Experiments and Analysis

The experimental results of the single index based on the DBLP_fourarea are shown in Table 1.

| Feature | F-Value | AUC  |
|---------|---------|------|
| CAR     | 0.0962  | 0.5491 |
| MCN     | 0.0831  | 0.5419 |
| MMCN    | 0.1095  | 0.5567 |
| ConfSim | \textbf{0.5299} | 0.4162 |
We can see that the performance of single CAR index is not that good. MMCN and ConfSim indexes take consider of domain knowledge, so their performance are better than CAR and MCN indexes according to AUC and F-value respectively, which shows that the domain knowledge is a significant factor for the improvement of link prediction methods. Experiments of MMCN index are better than CAR index, which illustrates the effectiveness of our algorithm. Table 2 shows experimental results of the feature vector based on combination of indexes, where CAR & DC represents that Features= \{CAR, DC\}.

Table 2. Results of DBLP fourarea based on the combination of indexes.

| Feature Vector                  | F-value | AUC  |
|---------------------------------|---------|------|
| CAR                             | 0.0962  | 0.5419|
| CAR&DC                          | 0.6493  | 0.4903|
| CAR&MNCN                        | 0.1831  | 0.5491|
| CAR&ConfSim                     | 0.4238  | 0.5097|
| CAR&DC&MNCN&ConfSim             | \textbf{0.6513} | \textbf{0.8839}|

As can be seen from Table 2, the feature vector which combines some indexes has improved performance of link prediction methods dramatically, which indicates that expanding information source is efficient. In addition, it can be seen that the feature vector combining with attributes of nodes has a certain improvement comparing with the single index, so we should utilize attributes of nodes and network context to enhance the accuracy of link prediction.

5. Conclusion

We propose a new method to predict links in social network. We calculate some indexes to represent topological information and attributes of nodes to construct feature vectors, and predict links based on XGB classifier. Experiments show that our method improves accuracy of link prediction effectively. In addition, components of the feature vector are easily calculated, so it can be applied to many types of networks and allows us pay more attention to the domain knowledge.

6. Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC Grant No. 61373051, No.61472159 and No.61572227), and the key support of e-commerce technology engineering laboratory innovation project of Jilin Province.

7. References

[1] Lü Linyuan and T Zhou. "Link prediction in complex networks: A survey." Physica A Statistical Mechanics & Its Applications390.6 (2011):1150-1170.
[2] Wang, Peng. et al. "Link prediction in social networks: the state-of-the-art." Science China Information Sciences 58.1(2015):1-38.
[3] Liben-Nowell, David and J Kleinberg. "The link prediction problem for social networks." Journal of the Association for Information Science & Technology 58.7(2007):1019–1031.
[4] Lin Dekang. "An Information-Theoretic Definition of Similarity." 15th Int. Conf. on Machine Learning Morgan Kaufmann Publishers Inc. 1998:296-304.
[5] Katz,Leo."A new status index derived from stoichiometric analysis. Psychometrika 18.1(1953):39-43.
[6] Freeman, Linton C. "Centrality in social networks conceptual clarification." Social Networks 1.3(1978):215-239.
[7] Sabidussi G "The centrality of a graph." Psychometrika31.4 (1966):581.
[8] Liu Haifeng, et al. "Hidden link prediction based on node centrality and weak ties." Epl 101.1(2013):18004.
[9] Li Yuhua, et al. "Research of Dynamic Link Prediction Method Based on Link Importance." Journal of Computer Research & Development (2011).
[10] Scellato, Salvatore, Anastasios Noulas and Cecilia Mascolo. "Exploiting place features in link prediction on location-based social networks." Proc. of the 17th ACM SIGKDD Int. Conf. on
Knowledge discovery and data mining. ACM, 2011.

[11] Hasan and a Mohammad. "Link Prediction using Supervised Learning." Proc. of Sdm Workshop on Link Analysis Counterterrorism & Security 30.9 (2006): 798-805.

[12] Cannistraci, Carlo Vittorio, G Alanislobato and T Ravasi. "Erratum: From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks." Sci Rep 3.4 (2013): 1613.

[13] Valverde-Rebaza J and A D A Lopes. "Structural link prediction using community information on Twitter." Int. Conf. on Computational Aspects of Social Networks IEEE, 2012: 132-137.

[14] Adamic, Lada A and E Adar. "Friends and neighbors on the Web." Social Networks 25.3 (2003): 211-230.

[15] Zhou Tao, L Lü and Y C Zhang. "Predicting missing links via local information." European Physical Journal B 71.4 (2009): 623-630.

[16] Chen Tianqi and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining. ACM, 2016.

[17] Babajide Mustapha, Ismail and Faisal Saeed. "Bioactive molecule prediction using extreme gradient boosting." Molecules 21.8 (2016): 983.