Research on Link Prediction Based on Similarity Index Algorithm

Zhaodi Xu
Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications Chongqing, China
834006386@qq.com

Abstract. For social network with no rights and no directions, link prediction has lower computational complexity and higher link prediction accuracy by using similarity metrics. The existing local information-based similarity metrics are mostly calculated by the common neighbors of the pair of nodes to be predicted when performing link prediction, but these methods do not consider the influence of the relationship between the common neighbors. Aiming at this problem, this paper proposes a new link prediction algorithm based on the existing local information-based similarity measurement method, which not only considers the number and degree of common neighbors between the pairs of nodes to be predicted, but also considers the relationship between common neighbors. In order to verify the effectiveness of the algorithm, the proposed algorithm is compared with the original algorithm and other methods based on the similarity measure of local information in this paper. The experimental results demonstrate the effectiveness of the algorithm.

1. Introduction
In the real world, relationships between entities can be abstracted into various complex networks such as online social networks, power networks, aviation networks and protein networks. Link prediction refers to the possibility of predicting two unconnected nodes that are hidden or may generate links in the future through the structural information or their information of the nodes. The problem of link prediction has great application value in real life. It can be used to recommend products to customers [1] and can be used for e-mail prediction [2]. It also can help to understand the micro-mechanism of network modeled and provide a theoretical basis for modeling social networks.

Link prediction for social networks is a hot topic which has being researched currently. The methods of link prediction mainly include similarity-based metrics, matrix-based factorization, and probability-based graph models. The methods of calculating the similarity value mainly include the similarity index based on the local information and the similarity index based on the path. Among them, the similarity index based on local information mainly includes common neighbor (CN) index [3], Jaccard index [4], resource allocation (RA) index [5], etc. Path-based similarity index includes local path (LP). Indicator [6], Katz indicator [7], etc. The methods based on matrix factorization include spectral transformation [8] and matrix decomposition [9-13], but the computational complexity of such methods is generally high. The methods based on probability map model include local probability map model [14], potential friend propagation model [15], and random walk [16] and so on.
For social networks with no rights and direction, the method for link prediction based on similarity measure has lower computational complexity and higher prediction accuracy, so people tend to choose method of calculating the similarity value to do link prediction of graphs which not be weighted and directed with the method of measuring the similarity of information or local paths. The existing local information-based link prediction method obtains the similarity value of the node to be predicted by calculating the common neighbor between the pairs of nodes to be predicted. But in the process, the common neighbor node is often considered as an independent node without considering the relationship between themselves. Aiming at this problem, this paper proposes a new link prediction algorithm based on the existing local information-based similarity measure method. In this algorithm, the relationship between public neighbors is considered in the calculation of similar values. And the accuracy of link prediction can be improved by using this algorithm.

2. Problem Description

2.1. Link prediction
A social network is defined as $G(V, E)$, where $V$ represents the set of nodes of the network and $E$ represents the set of relationships (directed or undirected) between the nodes in the network. The social network is divided into directed network and undirected network according to whether the edges have direction between the nodes or not, and is divided into a right network and an unprivileged network according to whether the edge exists values which are calculated by some information. Link prediction refers to a probability can be calculated through a link prediction algorithm that two points that are connected or may generate a relationship at the next moment. This article focuses on research on social networks which have no rights and direction. All the link prediction processes can be visually represented as Figure 1.

![Figure 1. The processes of link prediction](image)

2.2. The similarity indexes based on local information
The method based on the similarity measure considers that the higher similarity values of the two nodes, the more likely they would have a relationship at the next moment. There are several methods for the similarity index based on local information that are well known today:

1. Common Neighbors (CN) algorithm. This algorithm considers the number of common neighbors between two nodes to be predicted. If the number of common neighbors is higher, the more similar they are and the more likely the link between them will be. Assuming that the nodes to be predicted are $u_x$ and $u_y$, the similarity value of the nodes $u_x$ and $u_y$ can be calculated as shown in equation (1).

$$\text{CN}(u_{xy}) = |T(u_x) \cap T(u_y)|$$

(1)

Where $\text{CN}(u_{xy})$ represents the number of common neighbors of nodes $u_x$ and $u_y$, $T(u_x)$ represents the neighbor of node $u_x$, and $T(u_y)$ represents the neighbor node of node $u_y$.

2. Adamic-Adar (AA) algorithm. This algorithm is still considered the common neighbors between the nodes which would be predicted. But this algorithm not only considers the number of common
neighbors, but also considers the degree of the single neighbor node in the common neighbors. The main idea is that the neighbor nodes with small degree in the network have more influence on the similarity than the neighbor nodes with large degree. Assuming that the nodes to be predicted are \( u_x \) and \( u_y \), the similarity value of the nodes \( u_x \) and \( u_y \) can be calculated as shown in equation (2).

\[
AA(u_{xy}) = \frac{1}{d(z)} \sum_{z \in N(u_{xy})} \log d(z)
\]  

Where \( AA(u_{xy}) \) represents the similarity value of nodes \( u_x \) and \( u_y \), \( N(u_{xy}) \) represents the common neighbors of nodes \( u_x \) and \( u_y \), and \( d(z) \) represents the degree of public neighbors.

3. Jaccard algorithm. This algorithm calculates the ratio of the number of common neighbors between nodes to be predicted to the number of neighbors to be predicted. Assuming that the nodes to be predicted are \( u_x \) and \( u_y \), the similarity values of nodes \( u_x \) and \( u_y \) can be calculated as shown in equation (3).

\[
Jaccard(u_{xy}) = \frac{|T(u_x) \cap T(u_y)|}{|T(u_x) \cup T(u_y)|}
\]  

Where \( Jaccard(u_{xy}) \) represents the similarity value of nodes \( u_x \) and \( u_y \), \( T(u_x) \) represents the neighbor of node \( u_x \), and \( T(u_y) \) represents the neighbor node of node \( u_y \).

4. Resource Allocation (RA) algorithm. This algorithm considers the effect of common neighbor degree on similarity when calculating similar values. It is assumed that the node \( u_x \) passes the resource to \( u_y \) through the common neighbor nodes with \( u_y \), and the common neighbor would divide the resource equally to its neighbor node. The similarity value can be calculated as shown in formula (4).

\[
RA(u_{xy}) = \frac{1}{\text{degree}(z)} \sum_{z \in N(u_{xy})} \text{degree}(z)
\]  

Where \( RA(u_{xy}) \) represents the similarity value of nodes \( u_x \) and \( u_y \), \( N(u_{xy}) \) represents the common neighbors of nodes \( u_x \) and \( u_y \), and \( \text{degree}(z) \) represents the degree of common neighbors.

3. Algorithm Improvement

For two nodes that are not related to each other, the common neighbors between them is an important medium for the transfer relationship so that the calculation of the common neighbor is the main factor affecting the size of the similarity value. However, it is not enough just using the number and degree of common neighbors when do link prediction. For example, there are multiple common neighbors for the
nodes to be predicted \( u_x \) and \( u_y \), the relationship between \( u_x \) and \( u_y \) will be less than expected if only
the number of common neighbor nodes is considered, regardless of the common neighbor nodes. For a
real network, there is often a high correlation between neighbors that is these neighbors are connected
to each other. So that the relationship between the public neighbor nodes should be fully considered
when calculating similar values by the neighbor nodes. In order to solve the above problems, an
improved algorithm is proposed to consider the relationship between common neighbor nodes in the
calculation of similarity value based on RA which can improve the prediction accuracy.

As shown in Figure 2. The traditional RA algorithm considers that the common neighbors between
nodes A and B are independent and have no relationship. Therefore, when calculating the similarity
value, it is not necessary to consider

![Figure 2. The processes of link prediction](image)

4. Experimental design and results analysis

4.1. Datasets
In order to verify the effect of the proposed algorithm, the proposed algorithm and the other similar
methods based on local information calculations mentioned in this paper are compared on multiple
datasets. The datasets which are used in this experiment are the American Aviation Network (USAir),
the Protein Network (Yeast), the American Political Blog Network (PB), and the American College Football Network (ACF). The specific networks information is shown in Table 1.

| Datasets | Number of nodes | Number of edges | Average degree | Clustering coefficient |
|----------|-----------------|-----------------|-----------------|------------------------|
| USAir    | 332             | 2126            | 12.81           | 0.749                  |
| Yeast    | 2375            | 11693           | 9.85            | 0.388                  |
| PB       | 1202            | 16708           | 16.93           | 0.362                  |
| ACF      | 115             | 613             | 10.66           | 0.403                  |

4.2. Evaluation Index

The commonly used evaluation indexes for link prediction are P value and AUC value, but the size of P has a great relationship with the number of positive samples obtained so that is has great randomness. The value of AUC can measure the accuracy of the algorithm as a whole degree so that this paper chooses AUC as the evaluation index of the algorithm. The specific calculation method is shown in formula (5).

\[
AUC = \frac{n' + 0.5n''}{n}
\]  

This formula means: randomly select an edge from the test set and non-existent edge set in order. After independently selecting n times, \(n'\) represents how many times the score of the test edge is larger than that of the non-existent edge and \(n''\) represents how many times the score of test edge is equal to that of non-existent edge. Obviously, \(AUC \approx 0.5\) for random edge selection algorithm. Thus, the accuracy of the prediction algorithm depends on whether its AUC is significantly greater than 0.5 or not.

4.3. Experimental Results

The algorithm (NRA), CN algorithm, AA algorithm, Jaccard algorithm and RA algorithm are used to link prediction on the four kinds of datasets which are not considered node attributes and edge attributes. We use evaluation index AUC to judge the pros and cons of the algorithm and the comparison results are shown in Table 2.

| Dataset | CN       | AA       | Jaccard  | RA       | NRA      |
|---------|----------|----------|----------|----------|----------|
| USAir   | 0.8563   | 0.8662   | 0.8275   | 0.8793   | 0.8934   |
| Yeast   | 0.9124   | 0.9212   | 0.8893   | 0.9228   | 0.9321   |
| PB      | 0.8122   | 0.8235   | 0.8325   | 0.8427   | 0.8862   |
| ACF     | 0.9236   | 0.9039   | 0.9142   | 0.9241   | 0.9328   |

It can be seen from Table 2 that the proposed algorithm has better link prediction results than the other four algorithms for link prediction whether in the Air network dataset or in the other three datasets. Especially that the AUC on the dataset ACF reaches score with 0.9328. NRA is superior to other algorithms because the NRA algorithm fully considers the topology information between neighbor nodes.

In order to make the results look more intuitive, the AUC value of each link prediction algorithm on each dataset is represented by a line graph.
Figure 4. AUC on dataset named USAir

Figure 5. AUC on dataset named Yeast

Figure 6. AUC on dataset named PB

Figure 7. AUC on dataset named ACF
4.4. Analysis

It can be seen from Table 2 that the proposed algorithm has better link prediction results than the other four algorithms for link prediction whether in the Air network dataset or in the other three datasets. NRA is superior to other algorithms because the NRA algorithm fully considers the topology information between neighbor nodes.

As shown in figure 4, figure 5, figure 6 and figure 7, even though the original RA algorithm has a good prediction result, the result of the NRA algorithm is still better than the result of RA algorithm.

5. Conclusion

The NRA algorithm proposed in this paper not only considers the degree of public neighbor nodes, but also fully considers the relationship between common neighbor nodes. Experimental results show that the proposed algorithm has higher prediction accuracy.

The research in this paper is based on the social network without power and undirected. In fact, every node and edge have attributes, and there may have multiple relationships between nodes and nodes in real life. If multiple attributes are considered, there will be multiple edges, which are ignored in the study of this paper. In future research, more practical problems will be considered to improve the accuracy of link prediction.

Acknowledgments

This study was supported by Chongqing Key Laboratory of Computational intelligence at Chongqing University of Posts and Telecommunications. The authors also wish to thank anonymous reviewers and colleagues who have given some suggestions to improve this paper.

References

[1] Tu H, Tang X. Research on Product Recommendation Model Based on Label [J]. Modern Library and Information Technology, 2017, 1 (9): 28-39.
[2] Wang Q. Link prediction and threads in email networks [C] // International Conference on Data Science and Advanced Analytics. IEEE, 2015: 470-476.
[3] Newman M E J. Clustering and preferential attachment in growing networks [J]. Physical review E, 2001, 64 (2): 1-13.
[4] JACCARD P. Etude comparative de la distribution florale dans une portion des Alpes et des Jura [J]. Bull Soc Vaudoise Sci Nat, 1901, 37 (142): 547-579.
[5] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information [J]. European Physical Journal B, 2009, 71 (4): 623-630.
[6] Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks [J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2009, 80 (2): 046-122.
[7] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18 (1): 39-43.
[8] Kunegis J, Lommatzsch A. Learning spectral graph transformations for link prediction [C] // International Conference on Machine Learning. ACM, 2009: 561-568.
[9] Lee D. Learning the parts of objects with nonnegative matrix factorization [J]. Nature, 1999, 401 (6755): 788-788.
[10] Kuang D, Ding C, Park H. Symmetric nonnegative matrix factorization for graph clustering [C] // Proceedings of the 2012 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2012: 106-117.
[11] Dai C, Chen L, Li B, et al. Link Prediction in Multi-Relational Networks Based on Relational Similarity[J]. Information Sciences, 2017, 394(C):198-216.
[12] Zhu L, Guo D, Yin J, et al. Scalable temporal latent space inference for link prediction in dynamic social networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28 (10): 2765-2777.
[13] Kunegis J, Fliege J. Predicting directed links using nondiagonal matrix decompositions [C] //
[14] Wang C, Satuluri V, Parthasarathy S. Local Probabilistic Models for Link Prediction [C] // IEEE International Conference on Data Mining. Piscataway: IEEE, 2007: 322-331.

[15] Zhang J, Wang C, Yu P S, et al. Learning latent friendship propagation networks with interest awareness for link prediction [C] // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2013: 63-72.

[16] Yin Z, Gupta M, Weninger T, et al. Linkrec: a unified framework for link recommendation with user attributes and graph structure [C] // Proceedings of the 19th international conference on World wide web. New York: ACM, 2010: 1211-1212.