3C4-IOS-4a-3

Multi-class Link Prediction in Social Networks

Lankeshwara Munasinghe[*1] Ryutaro Ichise[*2]

[*1] Department of Informatics
The Graduate University for Advanced Studies, Tokyo, Japan
lankesh@nii.ac.jp

[*2] National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp

In this Internet era, online social network services play a vital role in decision making, providing numerous services to the users, etc. The sheer growth of such different social networks, which provide different services, has made it complicated for users to choose the most relevant social networking service(s) which meet their requirements. With these complications, link prediction in social networks has become a vital research topic. Most of the online social networks are composed of multi-class links which are related with different types of services. Thus, it is worthwhile to investigate the methods for predicting future links as well as the class or type of the links. In this paper, we presented time-related dependency method which is used for multi-class link prediction in heterogeneous social networks.

1. Introduction

Online social network services has become one of the most influential and key source of services providing, information/knowledge sharing and many other Internet based activities. Most of these social networks are represented as multi-relational networks. The users are linked with others via different types of links. Each class or type of links in network is related to a specific service or purpose. Thus, predicting the links which are more likely to happen as well as their types has become an important aspect of link prediction. Mainly, there are two types of heterogeneous social networks can be find. First, the same set of users are connected via different social networks [Dai 12]. Second, the same set of users are connected via different types of links within one social network [Davis 12]. Besides that, some research have tried to use similarities between different social networks with different users to predict links in one of them [Dong 12]. In the present study, we focused on first and second types of social networks. In both cases, our aim is to predict the future links between users with its type using the knowledge of other types of links that the users are already having. For example, a facebook user can comment on a post on his/her friend's wall. In other words, without being facebook friends a user can comment on third party's posts. Our aim is to find how likely them to become facebook friends after having linked via comments. In that case, the correlation between different types of links is a vital fact. Therefore, we focused on finding new methods to extract the knowledge from correlations between different types of links.

Multi-relational Social Networks can be modeled as a combination of a set of different social networks. The goal of multi-class link prediction is to find effective methods to predict links in one type of social network using the knowledge of the all social networks. Most of the previous research have utilized the social phenomenon such as cross network community formation[Cai 05, Tang 09] and triad formation[Dong 12] to study the evolution of heterogeneous social networks. In this paper, we investigated time-related patterns of triad formation in one type of links in the presence of other types of the links. We used the most recent time stamps of the links/interactions to understand the different patterns of triad formation. The time takes to appear the second type of link after the first type of link emerged/observed between a pair of nodes is a key indicator of the correlation between two types of links. One previous research has proposed triadic pattern based method for heterogeneous link prediction by considering type-related dependencies between different types of links [Davis 12] but has not considered time-related dependencies. Hence, we introduced a method to determine the time-related dependencies between different types of links and how this knowledge can used for heterogeneous link prediction.

Contact: Lankeshwara Munasinghe, The Graduate University for Advanced Studies, Tokyo, Japan. lankesh@nii.ac.jp

2. Supervised machine learning method for link prediction

Supervised machine learning methods are widely used in the past link prediction research [Munasinghe 13]. Hence, we used supervised machine learning to learn a model using the feature vectors of training data and the learned model used to predict the links in test data. The feature vectors are composed of some of the existing features which we used as the baseline features and the new feature introduced in this paper. We have shown the description of baseline features in the Table 1. Here, v_i, v_j and k denote nodes and Γ denotes the neighborhood of a node. Section 2.1 describes the new feature referred to as *time_related correlation label*.

Table 1: Baseline features

Feature	Description
Adamic/Adar	$\sum_{k \epsilon \Gamma(v_i) \bigcap \Gamma(v_j)} \frac{1}{log\lvert\Gamma(v_k)\rvert}$
Common neighbors	$\lvert\Gamma(v_i) \bigcap \Gamma(v_j)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(v_i) \bigcap \Gamma(j)\rvert}{\lvert\Gamma(v_i) \bigcup \Gamma(v_j)\rvert}$
Preferential attachment	$\lvert\Gamma(v_i)\rvert\lvert\Gamma(v_j)\rvert$

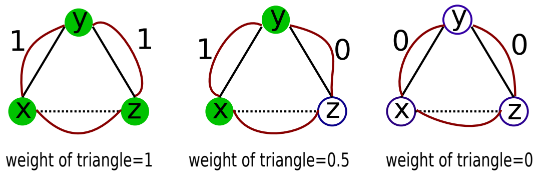

Figure 1: Triadic patterns

2.1 Time_related correlation label

The dependency/correlation between heterogeneous links depends on various factors. Time-related dependencies are one of the key factors which have not considered much in the previous research. The studies of general rules of social behaviors and our analytical results drew us to the assumption that if a correlation exists between type1 links and type2 links, the most recent link of them(say type2) should create within Δt time after the first one(type1) created(or observed). The time threshold, Δt, is defined as the weighted average of $|t_1 - t_2|$s of all type1 and type2 links. t_1 and t_2 are the most recent timestamps of the links of type1 and type2 respectively. The weighted average was calculated using the training data. The method assign higher weight as $|t_1 - t_2|$ decreases and lower weight as $|t_1 - t_2|$ increases. Once the system computed the weights, we can compute Δt which we used to assign a binary label for the edges, referred as *time_related correlation label*, in the following way.

$$time\_related\ correlation\ label = \begin{cases} 1 & \text{if } |t_1 - t_2| \leq \Delta t \\ 0 & \text{else} \end{cases}$$

The general intuition behind the *time_related correlation label* is that if $|t_1 - t_2| > \Delta t$, then there is no correlation between two types of links. This can be regarded as a measure of closeness between people. We observed that this assumption has a coincidence with the general social behaviors of the people.

The labeled edges used to identify different triad patterns in the given heterogeneous network. We have illustrated the possible triadic patterns in Figure 1. In this Figure, we are given the nodes x, y, z with edges (x, y) and (y, z). The brown and black color edges represent type1 and type2 links respectively. The task is to predict the type2 link (x, z). Here, we observed three possible triadic patterns. They are, both (x, y) and (y, z) are labeled with 1, one of them labeled with 1 and none of them labeled with 1. Now, we can interpret the given task as how likely the formation of each triangle. In order to do that, we assign a weight or score for each triangle based on the edge labels using a simple calculation.

$$weight_{\triangle} = \begin{cases} 1 & \text{if both edges labeled with 1} \\ 0.5 & \text{if one edge labeled with 1} \\ 0 & \text{if no edges labeled with 1} \end{cases}$$

The $weight_{\triangle}$ is used as a feature of a node pair in conjunction with supervised machine learning method for link prediction.

3. Conclusion

In this work, we investigated time-related dependency between different types of links in heterogeneous social networks and introduced a new method, referred as *time_related correlation label*, to label the multiple edges based on the difference between the most recent timestamps of the edges. The labeled edges used identify different triadic patterns in the network which give a very useful knowledge to predict future links. The experiments are being conducted to test the effectiveness of *time_related correlation label*.

Acknowledgments

We would like to be thankful to Nataliia Pobiedina of Vienna University of Technology for her valuable comments and sharing knowledge with us.

References

[Cai 05] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. *Knowledge Discovery in Databases*, pages 445–452, 2005.

[Tang 09] L. Tang, X. Wang, and H. Liu. Uncovering groups via heterogeneous interaction analysis. In *Proceedings of the Nineth IEEE International Conference on Data Mining*, pages 503 –512, 2009.

[Dong 12] Y. Dong, J. Tang, S. Wu, J. Tian, N.V. Chawla, J. Rao, and H. Cao Link prediction and recommendation across heterogeneous social networks. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 181–190, 2012.

[Davis 12] D. Davis, R. Lichtenwalter, and N.V. Chawla. Supervised methods for Multi-relational Link Prediction. *Social Network Analysis and Mining*, pages 1–15, 2012.

[Dai 12] B.T. Dai, F.C.T. Chua, and E.P. Lim. Structural analysis in multi-relational social networks. In *Proceedings of the SIAM International Conference on Data Mining*, 2012.

[Munasinghe 13] L. Munasinghe and R. Ichise. Link Prediction in Social Networks using Information Flow via Active Links. *IEICE Transactions*, Vol.E96-D, No.7, 2013.