Predicting Links and Their Building Time: A Path-based Approach

Manling Li, Yantao Jia, Yuanzhuo Wang, Zeya Zhao, Xueqi Cheng
CAS Key lab of Network Data Science & Technology, Institute of Computing Technology, CAS, Beijing, 100190, China
limanlingcs@gmail.com, jiayantao@ict.ac.cn, wangyuanzhuo@ict.ac.cn, zhaozeya@software.ict.ac.cn, cxq@ict.ac.cn

Abstract

Predicting links and their building time in a knowledge network has been extensively studied in recent years. Most structure-based predictive methods consider structures and the time information of edges separately, which fail to characterize the correlation between them. In this paper, we propose a structure called the Time-Difference-Labeled Path, and a link prediction method (TDLP). Experiments show that TDLP outperforms the state-of-the-art methods.

Introduction

Predicting links and their building time in a knowledge network, i.e., a network with multiple typed vertices and time-labeled edges, is important to detect the evolution of a dynamic network and has been paid much attention. Actually, we may be more interested in “Will two authors co-write a paper within 5 years?” than “Will two authors co-write a paper?”. Nevertheless, the traditional structure-based methods, e.g., the path ranking algorithm (PRA) (Lao et al., 2012), used the paths without the time information to predict the existence of links rather than the building time of links. Recently, a meta-path based predictive method GLM (Sun et al., 2012) was proposed to predict the building time of links, and was proved to be the state-of-the-art predictive model. However, it considered structures and the time information separately, but failed to integrate the time information of links into the path features.Thus, it was unable to distinguish paths with different timestamps of links, which is indispensable because a link is more likely to recur in the future if it has appeared recently (Rossetti et al., 2011). Consequently, how to combine structures and the time information to promote the performance of temporal link prediction is imperative. To address this issue, we propose a Time-Difference-Labeled Path based method (TDLP for short) by modeling the time-involving path. The contribution of TDLP is to integrate the time information into the path features, and propose a predictive method superior to the state-of-the-art methods.

Links and Building Time Prediction

In this study, we simply model the knowledge network as a time-involving graph $G = (V, E, R, T)$, where V denotes the set of vertices. E denotes the set of edges (v_i, v_j, r_k), $v_i, v_j \in V$, $r_k \in R$, where R is the set of edge type. And T is the set of building time of edges. We will firstly define the time-difference-labeled path, and then establish TDLP method.

Time-Difference-Labeled Path

Time-difference-labeled path is a path with all edges labeled with time difference. More formally, given two vertices v_0, v_l, and edge type r, suppose that the building time of the edge (v_0, v_l, r) is predicted as t^*. A time-difference-labeled path denoted by $P_{v_0v_l}$ with length l is defined as

$$P_{v_0v_l} = (r_1, \Delta t_1)(r_2, \Delta t_2)\ldots(r_i, \Delta t_i)\ldots(r_l, \Delta t_l),$$

where t_i donates the building time of edge (v_{i-1}, v_i, r_i), and $\Delta t_i = t^* - t_i$ for $i = 1, 2, \cdots, l$.

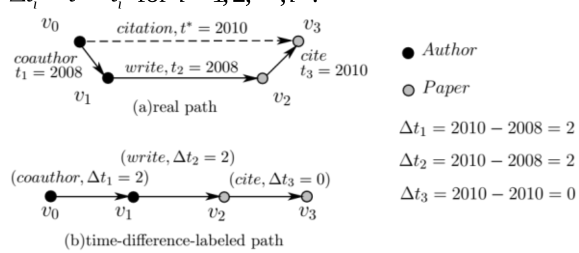

Figure 1 Process of generating a time-difference-labeled path $P_{v_0v_3} = (coauthor, 2)(write, 2)(cite, 0)$ *between* v_0 *and* v_3

For example, given v_0, v_3 where $(v_0, v_3, citation)$ is predicted to build at $t^* = 2010$, we firstly find one path $v_0 \rightarrow v_1 \rightarrow v_2 \rightarrow v_3$ in the knowledge network as shown in Figure 1(a). Then, we calculate the time difference Δt_i between the predicted time t^* and the building time t_i of three edges, e.g., $\Delta t_1 = 2010 - 2008 = 2$. Finally, we obtain the time-difference-labeled path $P_{v_0v_3}$ in Figure 1(b).

TDLP Method

TDLP method is conducted in a supervised setting. Firstly, for each pair of vertices in the training data, we construct all possible time-difference-labeled paths with different lengths l by the well-known breadth-first traversal, and form the set $P=\{P_{v_0v_l}^{(1)}, P_{v_0v_l}^{(2)}, \cdots, P_{v_0v_l}^{(n)}\}$, where n is the number of different paths. Secondly, we use P as features and learn their weights by maximum likelihood estimation. During the predictive process, to predict the building time of edge (v_0, v_l, r), we rank all potential timestamps $\{t^*\}$, in terms of the scores obtained by combining the different weighted path features between v_0 and v_l with different lengths.

Specifically, for a time-difference-labeled path $P_{v_0v_l}^{(i)}$ for $i=1,2,\cdots,n$ and the predicted time t^*, we define the score of path $P_{v_0v_l}^{(i)}$ as $S\left(P_{v_0v_l}^{(i)}\right)$ recursively as follows.

If $l=0$, then $P_{v_0v_l}^{(i)}$ is an empty path, and set $S\left(P_{v_0v_l}^{(i)}\right)=1$. If $l>0$, let $B_{v_l}=\{e \mid e \xrightarrow{r_l, \Delta t_l} v_l\}$ be the set of the neighbors of v_l, whose edge type with v_l is r_l and the label of time difference is Δt_l. Then we define that

$$S\left(P_{v_0v_l}^{(i)}\right)=\sum_{e' \in B_{v_l}} S\left(P_{v_0e'}^{(i)}\right) \cdot Pr(v_l \mid e', r_l, \Delta t_l),$$

Where $Pr(v_l \mid e', r_l, \Delta t_l)$ is the probability of reaching v_l from e' with a one-step random walk labeled as r_l and Δt_l. Namely,

$$Pr(v_l \mid e', r_l, \Delta t_l)=\sigma(v_l, e' \mid r_l, \Delta t_l) / \sigma(v_l, * \mid r_l, \Delta t_l)$$

where $\sigma(v_l, e' \mid r_l, \Delta t_l)$ indicates whether there exists a link typed r_l from e' to v_l with Δt_l, and $\sigma(v_l, * \mid r_l, \Delta t_l)$ calculates the number of links typed r_l from any node to v_l with Δt_l. By linearly combining the feature values of different labeled paths $P_{v_0v_l}^{(i)}$ with different lengths, we obtain the accumulated score $Score(t^*)$ of time t^* by

$$Score(t^*)=\sum_{i=1}^{n} S\left(P_{v_0v_l}^{(i)}\right) \cdot \lambda_i,$$

where λ_i is the weight of the feature score $S\left(P_{v_0v_l}^{(i)}\right)$. We follow the way of PRA to determine λ_i by maximum likelihood estimation. More detail can be referred to (Lao et.al., 2012). Notice that, if $Score(t^*)$ is larger than the threshold d, the predictive building time is t^*. Otherwise, the output is set to be ∞, which means that (v_0, v_l, r) will not exist in the future.

Experiment

The experiments are carried out on ArnetMiner[1], an academic network, consisting of four types of vertices, i.e., Author, Paper, Venue and Key Word. We select top 5000 active authors who published more than 5 papers between 2000 and 2013. In this paper, we concentrate on predicting the building time of four types of links, namely, *coauthor* between Authors, *citation* between Author and Paper, *mention* between Author and Key Word, *contribute* between Author and Venue. For each type r we construct the training data from 2000 to 2008, and carry out 5-fold cross-validation to learn the weight of paths. In the experiments, we set path length $l=1,2,3,4$, and the best predictive accuracy achieves when the threshold d is set to 0.6. To evaluate the performance of temporal link prediction, we test the methods on the data from 2009 to 2013 by two stages:

1) Predicting whether an edge will build in the future. Here, we employ Accuracy, and choose three methods PRA, GLM_exp and GLM_geo as baselines listed in Table 1.

2) Predicting the building time of upcoming edges. Here, we employ MAE and RSME, and choose methods GLM_exp and GLM_geo as baselines listed in Table 2.

Table 1. Comparison of Accuracy(%)

Link type	*GLM_exp*	*GLM_geo*	*PRA*	*TDLP*
coauthor	64.13	60.02	68.24	**70.31**
citation	67.19	69.93	72.47	**74.78**
focus	67.25	63.57	70.92	**73.60**
contribute	70.92	67.48	74.16	**75.87**

Table 2. Comparison of MAE and RSME

Link Type	*GLM_exp*		*GLM_geo*		*TDLP*	
	MAE	*RSME*	*MAE*	*RSME*	*MAE*	*RSME*
coauthor	4.20	24.97	2.64	15.22	**1.22**	**8.09**
citation	4.31	27.68	3.61	22.74	**1.44**	**9.39**
mention	5.10	32.24	3.19	16.23	**1.81**	**8.91**
contribute	4.06	23.93	3.77	21.17	**1.17**	**6.52**

It can be seen that: 1) TDLP is more accurate in predicting link existence from Table 1; 2) and TDLP obtains the lowest MAE and RSME from Table 2. It is unsurprising since that TDLP regards the time information and the path information as a unified feature and models their interplay in an intrinsic way. On the contrary, PRA ignores time information, and the other two baselines consider the time information and topological information separately.

Conclusion

In this paper, we proposed TDLP method for predicting links and their building time in a knowledge network, which combines the time and structural information into a unified setting, and experiments demonstrate the effectiveness of the proposed method.

Acknowledgments

This work is supported by National Grand Fundamental Research 973 Program of China (No. 2013CB329602, 2014CB340401), National Natural Science Foundation of China (No. 61173008, 61402442, 61572469, 61572473, 61303244), Beijing nova program(No. Z121101002512063), and Beijing Natural Science Foundation (No. 4154086).

References

Lao N, Subramanya A, Pereira F, and Cohen W. W. 2012. Reading the web with learned syntactic-semantic inference rules. In *Proc. EMNL*, 1017-1026.

Sun, Y., Han, J., Aggarwal, C. C and Chawla, N. V. 2012. When will it happen?: relationship prediction in heterogeneous information networks. In *Proc. WSDM*, 663-672.

Rossetti G, Berlingerio M, Giannotti F. 2011. Scalable link prediction on multidimensional networks. In *Proc. ICDMW*, 979-986.

[1] https://aminer.org/