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, jayantao@ict.ac.cn, wangyuanzhuo@ict.ac.cn, zhaoyeza@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 \( \{v_i, v_j, r_i\} \), \( v_i, v_j \in V \), \( r_i \in R \), 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 \( v_a, v_b \), and edge type \( r \), suppose that the building time of the edge \( (v_a, v_b, r) \) is predicted as \( t^* \). A time-difference-labeled path denoted by \( P_{v_a v_b} \) with length \( l \) is defined as
\[
P_{v_a v_b} = (r_i, \Delta t_i)(r_2, \Delta t_2)\ldots(r_l, \Delta t_l),
\]
where \( t_i \) donates the building time of edge \( (v_{i-1}, v_i, r_i) \), and \( \Delta t_i = t^* - t_i \) for \( i = 1, 2, \ldots, l \).

Figure 1: Process of generating a time-difference-labeled path

For example, given \( v_0, v_3 \) where \( (v_0, v_3) \) 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 \( \Delta t \) between the predicted time \( t^* \) and the building time \( t_i \) of three edges, e.g., \( \Delta t_1 = 2010-2008 = 2 \), \( \Delta t_2 = 2010-2008 = 2 \), \( \Delta t_3 = 2010-2008 = 0 \). Finally, we obtain the time-difference-labeled path \( P_{v_0 v_3} \) 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 $P = \{P_{v_i}^{(1)}, P_{v_i}^{(2)}, \ldots, P_{v_i}^{(n)}\}$, 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 $(v_i, v_j, r)$, we rank all potential timestamps $\{r'\}$, in terms of the scores obtained by combining the different weighted path features between $v_i$ and $v_j$ with different lengths.

Specifically, for a time-difference-labeled path $P_{v_i}^{(i)}$ for $i = 1, 2, \ldots, n$ and the predicted time $r'$, we define the score of path $P_{v_i}^{(i)}$ as $S(P_{v_i}^{(i)})$ recursively as follows.

If $l = 0$, then $P_{v_i}^{(i)}$ is an empty path, and set $S(P_{v_i}^{(i)}) = 1$. If $l > 0$, let $B_{v_i} = \{e \mid e \rightarrow (v_i, r_i, \Delta t_i)\}$ be the set of the neighbors of $v_i$, whose edge type with $v_i$ is $r_i$ and the label of time difference is $\Delta t_i$. Then we define that

$$S(P_{v_i}^{(i)}) = \sum_{e \in B_{v_i}} S(P_{v_i}^{(i)}) \cdot P_r (e \mid e', r_i, \Delta t_i),$$

where $Pr(e \mid e', r_i, \Delta t_i)$ is the probability of reaching $v_i$ from $e'$ with a one-step random walk labeled as $r_i$ and $\Delta t_i$. Namely,

$$Pr(e \mid e', r_i, \Delta t_i) = \sigma(v_i, e', r_i, \Delta t_i) / \sigma(v_i, e', \Delta t_i),$$

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

$$Score(t') = \sum_{v_i} S(P_{v_i}) \cdot \lambda,$$

where $\lambda$ is the weight of the feature score $S(P_{v_i})$. We follow the way of PRA to determine $\lambda$ 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 $\infty$, which means that $(v_i, v_j, r)$ will not exist 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>
<thead>
<tr>
<th>Link Type</th>
<th>GLM_exp</th>
<th>GLM_geo</th>
<th>PRA</th>
<th>TDLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>coauthor</td>
<td>64.13</td>
<td>60.02</td>
<td>68.24</td>
<td>70.31</td>
</tr>
<tr>
<td>citation</td>
<td>67.19</td>
<td>69.93</td>
<td>72.47</td>
<td>74.78</td>
</tr>
<tr>
<td>focus</td>
<td>67.25</td>
<td>63.57</td>
<td>70.92</td>
<td>73.60</td>
</tr>
<tr>
<td>contribute</td>
<td>70.92</td>
<td>67.48</td>
<td>74.16</td>
<td>75.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link Type</th>
<th>MAE</th>
<th>RSME</th>
<th>MAE</th>
<th>RSME</th>
</tr>
</thead>
<tbody>
<tr>
<td>coauthor</td>
<td>4.20</td>
<td>2.64</td>
<td>15.22</td>
<td>1.22</td>
</tr>
<tr>
<td>citation</td>
<td>4.31</td>
<td>3.61</td>
<td>22.74</td>
<td>1.44</td>
</tr>
<tr>
<td>mention</td>
<td>5.10</td>
<td>3.19</td>
<td>16.23</td>
<td>1.81</td>
</tr>
<tr>
<td>contribute</td>
<td>4.06</td>
<td>3.77</td>
<td>21.17</td>
<td>1.17</td>
</tr>
</tbody>
</table>

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**

