

基于 GPS 轨迹数据的拥堵路段预测

林树宽, 于伶姿, 乔建忠, 张百合
(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 基于真实的 GPS 轨迹数据, 对城市拥堵路段进行预测. 在此过程中, 摒弃传统的基于交通流预测和拥堵识别的方法, 提出一种新的基于拥堵向量和拥堵转移矩阵的拥堵路段预测方法. 该方法同时考虑路段拥堵的时间周期性和时空相关性, 通过对出租车 GPS 轨迹数据进行挖掘和训练, 建立拥堵向量和拥堵转移矩阵, 实现对拥堵路段的预测. 真实数据集上的实验验证了所提的拥堵路段预测方法的有效性.

关 键 词: GPS 轨迹数据; 路段间时空因果关系; 拥堵转移概率; 拥堵转移矩阵; 拥堵路段预测
中图分类号: TP 391 **文献标志码:** A **文章编号:** 1005-3026(2015)11-1530-05

The Congestion Road Segment Prediction Based on GPS Trajectory Data

LIN Shu-kuan, YU Ling-zi, QIAO Jian-zhong, ZHANG Bai-he

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: LIN Shu-kuan, professor, E-mail: linshukuan@ise.neu.edu.cn)

Abstract: Congestion road segments over real GPS trajectory data was predicted. The traditional methods were ostracized based on traffic flow prediction and congestion identification, and a novel method was proposed based on the congestion vector and the congestion transition matrix. The congestion vector and the congestion transition matrix were established by mining and training taxi GPS trajectory data, resulting in the implementation of the prediction of traffic congestion. In the course of prediction, time periodicity and spatial-temporal correlation of road segment congestion were both considered. The experiments on real data show the effectiveness of the congestion road segment prediction method proposed.

Key words: GPS trajectory data; spatial-temporal causal relationship between road segments; congestion transition probability; congestion transition matrix; congestion road segment prediction

近年来,城市机动车数量增加迅速,道路资源建设相对滞后,导致城市交通拥堵日益严重,拥堵路段预测成为建设智能城市和交通不可缺少的组成部分.越来越多的学者开始关注和研究基于交通数据的拥堵路段预测方法和技术^[1-4].现有的相关研究多基于道路上固定安装的传感器采集的数据,且多集中在交通拥堵路段的识别、挖掘和交通流预测方面^[5-9].

近年来,车载 GPS 发展迅速,产生大量的时空数据,可用于进行拥堵路段预测.本文基于出租车 GPS 轨迹数据,提出了一种新颖的考虑路段拥堵时间周期性和时空因果关系的、基于拥堵向量

和拥堵转移矩阵的拥堵路段预测方法.

1 拥堵转移矩阵的建立

本文同时考虑拥堵路段具有的时间周期性和时空相关性,提出了基于拥堵向量和拥堵转移矩阵的拥堵路段预测方法.为了建立拥堵转移矩阵,需要对历史 GPS 轨迹数据进行挖掘和训练.车流与路段拥堵情况在工作日(周一至周五)和休息日(周六和周日)呈现不同的规律,为此,本文将工作日和休息日的历史轨迹数据进行训练,并将工作日或休息日中的一天看作一个周期.路段拥

收稿日期: 2014-10-23

基金项目: 国家自然科学基金资助项目(61272177).

作者简介: 林树宽(1966-),女,吉林长春人,东北大学教授;乔建忠(1964-),男,辽宁兴城人,东北大学教授,博士生导师.

堵具有时间周期性,虽然同一周期的不同时间段呈现不同的拥堵状况,但不同周期的相同时间段却具有相似的拥堵情况,因此,本文将每个周期划分为具有固定时间长度的时间段,称为时间框架.历史轨迹数据的训练过程就是通过挖掘各个周期和时间框架内的历史拥堵路段,建立各时间框架的拥堵转移矩阵,为未来时间框架内的拥堵路段预测提供支持.

1.1 拥堵路段挖掘

为了在训练阶段建立拥堵转移矩阵,需要挖掘出每个周期和时间框架内的拥堵路段.由 GPS 采集的轨迹数据是由轨迹点构成的序列,轨迹模式为 $tr = (q_1, q_2, \dots, q_n)$,其中每个轨迹点 q_i ($1 \leq i \leq n$) 是一个二元组,即 $q_i = (p_i, t_i)$,这里 p_i 是位置信息, t_i 是时间信息.在挖掘拥堵路段之前,需要对历史轨迹数据进行预处理,即从轨迹数据中提取路段集合和经过每条路段的轨迹片段集合.本文基于文献[10]中的方法从轨迹数据中挖掘十字路口,在此基础上,形成由相邻十字路口构成的路段集合以及经过每条路段的轨迹片段集合,具体过程如算法 1 所示.

算法 1 路段集合和路段上的轨迹片段集合提取算法.

输入:历史 GPS 轨迹集合 T ,十字路口集合 I .

输出:路段集合 RS ,每条路段上的轨迹片段集合 TS 和后继路段集合 $Successors$.

```

1.  $RS = \emptyset$ ;
2. for 每条轨迹  $tr \in T$  do
3.   取轨迹  $tr$  的第一个十字路口点  $q_1$ ;
4.   for  $q_1$  之后的邻接轨迹点  $q_2$  do
5.     if  $q_2 \in I$  then
6.       if (由  $q_1$  和  $q_2$  构成的路段  $r \in RS$ )
7.          $r.TS = r.TS \cup (q_1, q_2, tr)$ ;
8.       else
9.          $r.TS = \emptyset$ ;  $r.Successors = \emptyset$ ;
10.         $RS = RS \cup r$ ;
11.      if ( $q_1$  不是  $tr$  的首个十字路口点)
12.         $r0.Successors = r0.Successors \cup r$ ;
13.       $q_2 \rightarrow q_1$ ;  $r \rightarrow r0$ .
```

算法 1 遍历每条轨迹的每个轨迹点,将相邻十字路口点 q_1 和 q_2 构成的路段 r 加入路段集合 RS ,并形成经过每个路段 r 的轨迹片段集合 TS

和路段 r 的后继路段集合 $Successors$, TS 中的元素为由 (q_1, q_2, tr) 构成的三元组.

在算法 1 提取的路段集合和路段上的轨迹片段集合的基础上,对各周期和时间框架内的拥堵路段进行挖掘,如算法 2 所示.首先,提取每个路段上的轨迹片段所属的周期和时间框架,并形成每个路段在各周期和时间框架的轨迹集合(行 2~6);其次,对于每个路段,计算各周期和时间框架内经过路段的轨迹数量,并根据路段饱和度判断路段在相应的周期和时间框架内是否拥堵,从而形成各周期和时间框架的拥堵路段集合(行 7~15).

算法 2 各周期和时间框架内的拥堵路段挖掘算法.

输入:路段集合 RS ,每条路段上的轨迹片段集合 TS ,每个路段饱和度 $capacity$.

输出:各周期和时间框架的拥堵路段集合 $congestionResult$.

```

1.  $congestionResult = \emptyset$ ;
2. for 每个路段  $r \in RS$  do
3.   for 每个轨迹片段  $(q_1, q_2, tr) \in r.TS$  do
4.     基于  $q_1, q_2$  的时间信息提取轨迹片段所属周期  $\rightarrow d$ ;
5.     基于  $q_1, q_2$  的时间信息提取轨迹片段所属时间框架  $\rightarrow f$ ;
6.      $tr \rightarrow$  路段  $r$  的周期为  $d$ 、时间框架为  $f$  的集合  $S(r, d, f)$  中;
7.   for 每个周期  $d$  do
8.     for 每个时间框架  $f$  do
9.        $R^{(f \in d)} = \emptyset$ ;
10.    for 每条路段  $r \in RS$  do
11.      计算集合  $S(r, d, f)$  的大小  $\rightarrow num$ 
12.      if  $num(r, d, f) > capacity(r)$ 
13.         $r.d.f.isCongestion = true$ ;
14.         $R^{(f \in d)} = R^{(f \in d)} \cup r$ ;
15.      else  $r.d.f.isCongestion = false$ ;
16.     $congestionResult = congestionResult \cup R^{(f \in d)}$ .
```

1.2 拥堵转移矩阵建立

1.1 节考虑了路段拥堵的时间周期性,从 GPS 轨迹数据中挖掘各周期和时间框架内的历史拥堵路段.从日常生活观察中可以发现,路段拥堵不仅具有时间周期性,还具有很强的时空相关性.所谓路段拥堵的时空相关性是指某些拥堵路

段在时间、空间上具有一定的关联,一条路段在一个时间框架内拥堵可能造成与它相邻的路段在下一个时间框架内也发生拥堵.本节将基于 1.1 节得到的拥堵路段集合,挖掘路段之间的时空因果关系,从而建立拥堵转移矩阵.本文假设拥堵路段之间的时空相关性具有一阶马尔科夫性:即当前路段是否拥堵只与和它最相邻的路段相关,并且只与前一个时间框架相关.为了更清楚地描述拥堵转移矩阵的建立过程,先给出相关定义.

定义 1 (路段的时空因果关系) 对于时间框架 f 内的路段 $r_i = (q_i^s, q_i^e)$ 和时间框架 f' 内的路段 $r_j = (q_j^s, q_j^e)$, 若满足: $(q_i^e = q_j^s)$ 且 $f' = f + 1$, 则称两路段 r_i 和 r_j 关于时间框架 f 具有时空因果关系, 记为 $r_i \xrightarrow{(f)} r_j$.

定义 1 中的 q_i^s 和 q_i^e 分别为路段 r_i 的两个端点.

定义 2 (路段的拥堵因果关系) 对于满足 $r_i \xrightarrow{(f)} r_j$ 关系的路段 r_i, r_j , 若 $r_i \in R^{(f \in d)}$ 且 $r_j \in R^{((f+1) \in d)}$, 则称路段 r_i, r_j 在周期 d 关于时间框架 f 具有拥堵因果关系, 记为 $r_i \xRightarrow{(f \in d)} r_j$; 若 $r_i \in R^{(f \in d)}$, 但 $r_j \notin R^{((f+1) \in d)}$, 则称路段 r_i, r_j 在周期 d 关于时间框架 f 不具有拥堵因果关系, 记为 $\neg r_i \xRightarrow{(f \in d)} r_j$.

定义 2 中的 $R^{(f \in d)}$ 为算法 2 挖掘出的周期 d 中时间框架 f 内的拥堵路段集合, $r_i \in R^{(f \in d)}$ 表示路段 r_i 在周期 d 中的时间框架 f 内发生拥堵.

定义 3 (路段拥堵因果关系矩阵) 周期 d 的时间框架 f 中的拥堵因果关系矩阵是一个 $n \times n$ 矩阵, 记为 $M^{(f \in d)}$, n 为路段集合 RS 中的路段总数. 矩阵元素 $M_{ij}^{(f \in d)}$ ($n \geq i, j \geq 1$) 共有 3 种取值, 即

$$\left. \begin{aligned} M_{ij}^{(f \in d)} &= 1, \text{ 若 } r_i \xRightarrow{(f \in d)} r_j; \\ M_{ij}^{(f \in d)} &= -1, \text{ 若 } \neg r_i \xRightarrow{(f \in d)} r_j; \\ M_{ij}^{(f \in d)} &= 0, \text{ 其他情况.} \end{aligned} \right\} \quad (1)$$

各周期的每个时间框架内的路段拥堵因果关系矩阵可依据定义 3, 基于算法 2 得到的拥堵路段集合生成, 具体生成过程如算法 3 所示.

算法 3 路段拥堵因果关系矩阵生成.

输入: 各周期和时间框架的拥堵路段集合 congestionResult.

输出: 所有周期的各时间框架中的路段拥堵因果关系矩阵集合 MS.

1. $MS = \emptyset$;
2. **for** 每个周期 d **do**
3. **for** 每个时间框架 $f \in d / \text{lasttimeframe}$ **do**

4. 将矩阵 $M^{(f \in d)}$ 的所有元素初始化为 0;

5. **for** 每个路段 $r_i \in R^{(f \in d)}$

6. **for** 每个路段 $r_j \in r_i$. Successors

do

7. **if** $r_j \in R^{((f+1) \in d)}$ **then**

8. $M_{ij}^{(f \in d)} = 1$;

9. **else**

10. $M_{ij}^{(f \in d)} = -1$;

11. $MS = MS \cup M^{(f \in d)}$.

定义 4 (拥堵转移概率) 路段 r_i 与路段 r_j 之间关于时间框架 f 的拥堵转移概率 $p_{ij}^{(f)}$ 定义为路段 r_i 在时间框架 f 内拥堵的情况下路段 r_j 在时间框架 $f+1$ 内发生拥堵的概率, 即

$$p_{ij}^{(f)} = P(r_j \in R^{(f+1)} | r_i \in R^{(f)}) . \quad (2)$$

根据定义 4, 利用算法 3 生成的路段拥堵因果关系矩阵, 路段 r_i 与路段 r_j 之间关于时间框架 f 的拥堵转移概率可计算如下:

$$p_{ij}^{(f)} = \sum_{d=1 \wedge M_{ij}^{(f \in d)} > 0}^N M_{ij}^{(f \in d)} / \sum_{d=1}^N |M_{ij}^{(f \in d)}| . \quad (3)$$

其中: N 为周期总数; 分子可计算出在所有周期的时间框架 f 中路段 r_i 发生拥堵的情况下路段 r_j 发生拥堵的计数; 分母可计算出在所有周期的时间框架 f 中路段 r_i 发生拥堵的情况下, 路段 r_j 发生拥堵和不发生拥堵的总计数.

定义 5 (拥堵转移矩阵) 关于时间框架 f 的拥堵转移矩阵 $P^{(f)}$ 是一个 $n \times n$ 矩阵, n 为路段集合 RS 中的路段总数, 元素 $P_{ij}^{(f)}$ 为路段 r_i 与路段 r_j 之间关于时间框架 f 的拥堵转移概率.

根据定义 5, 各时间框架所对应的拥堵转移矩阵可基于式 (3) 计算的拥堵转移概率生成, 从而为进行拥堵路段预测提供支持.

2 拥堵路段预测

拥堵路段预测包含两部分: 第一部分为当前拥堵路段计算, 通过统计当前时间框架各路段上的交通流, 计算当前时间框架内的拥堵路段, 并表示为向量形式; 第二部分则对未来时间框架的拥堵路段进行预测.

定义 6 (拥堵向量) 时间框架 f 内的拥堵向量 $V^{(f)}$ 是一个 n 元向量, 即 $V^{(f)} = (p_1^{(f)}, p_2^{(f)}, \dots, p_n^{(f)})$, 其中, n 为路段集合 RS 中的路段总数, 元素 $p_i^{(f)}$ ($1 \leq i \leq n$) 为路段 r_i 在时间框架 f 内拥堵的概率, 即

$$p_i^{(f)} = P(r_i \in R^{(f)}) . \quad (4)$$

当前时间框架的拥堵向量可以基于轨迹数据,统计当前时间框架内各路段的车流量,如果车流量大于饱和度^[9],表示路段发生拥堵,则将路段对应的拥堵向量元素设为 1,否则设为 0.

在 1.2 节得出的拥堵转移矩阵和以上得出的拥堵向量的基础上,通过计算未来时间框架的拥堵向量,可以对未来时间框架的拥堵路段进行预测,具体描述见定理 1.

定理 1 设当前时间框架 f 内的拥堵向量和拥堵转移矩阵分别为 $V^{(f)}$ 和 $P^{(f)}$, 则时间框架 $f+1$ 内的路段拥堵向量 $V^{(f+1)} = V^{(f)} \cdot P^{(f)}$.

证明 (略).

按照定理 1,基于当前时间框架的拥堵向量和拥堵转移矩阵,可以计算未来时间框架的拥堵向量.

给定拥堵概率阈值 β ,若拥堵向量中的拥堵概率大于阈值,表示相应路段拥堵,从而对拥堵路段进行预测.

3 实验分析

实验采用的数据集为真实数据集 T-Drive (<http://research.microsoft.com/apps/pubs/?id=152883>). 所有实验程序的开发环境为 3.16GHz Intel Core(TM2) Duo CPU,4GB 内存,操作系统为 Windows XP,编程语言为 C++.

实验将本文提出的基于拥堵转移矩阵的预测方法 PBTM (prediction based on transferring matrix) 与文献[9]中基于交通流转移概率的方法 PBP (prediction based on probability) 进行比较. 实验中涉及到的评价指标(准确率、召回率)如式(9)和式(10)所示:

$$\text{precision} = \frac{|V_{\text{rgt}}|}{|V_r|}, \quad (9)$$

$$\text{recall} = \frac{|V_{\text{rgt}}|}{|V_{\text{gt}}|}. \quad (10)$$

其中: $|V_{\text{rgt}}|$ 是预测得到的并且真实拥堵的路段数量; $|V_r|$ 是预测得到的拥堵路段数量; $|V_{\text{gt}}|$ 是真实的拥堵路段数量.

图 1 给出了两种方法的预测准确率随时间框架长度的变化情况.

从图 1 可以看出,随着时间框架长度的增加, PBTM 和 PBP 的预测准确率都降低. 但 PBTM 的准确率高于 PBP, 因为 PBTM 更符合路段拥堵的特点, 不仅考虑了路段拥堵的时间周期性, 而且考

虑了拥堵路段间的时空因果关系.

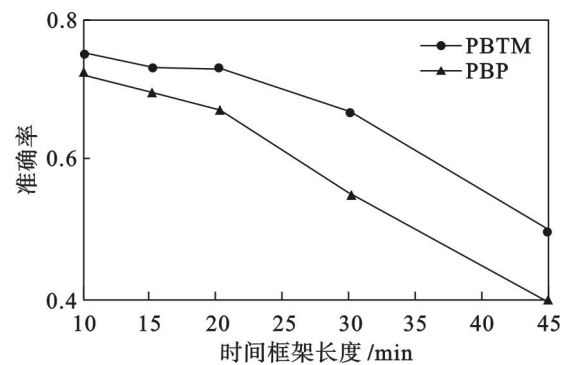


图 1 时间框架长度对预测准确率的影响

Fig. 1 The effect of length of time frame on precision

图 2 给出了两种方法的预测准确率随训练数据集大小的变化情况, 其中时间框架长度为 15min. 从图 2 可以看出, 随着训练数据集的增大, PBTM 和 PBP 的预测准确率都增大, 具有较好的扩展性. 对于 PBTM, 训练数据集越大, 包含的周期数量就越多, 拥堵路段间的时空因果关系挖掘的越完全, 拥堵转移矩阵也就越准确. 对于相同的拥堵向量, 预测得到正确的拥堵路段越多, 准确率越高.

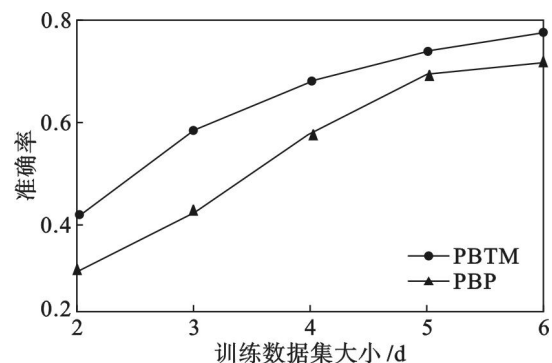


图 2 训练数据集大小对预测准确率的影响

Fig. 2 The effect of size of training dataset on precision

图 3 显示了阈值 β 的变化对预测准确率和召回率的影响. 从图 3 可以看出, 随着阈值 β 的增

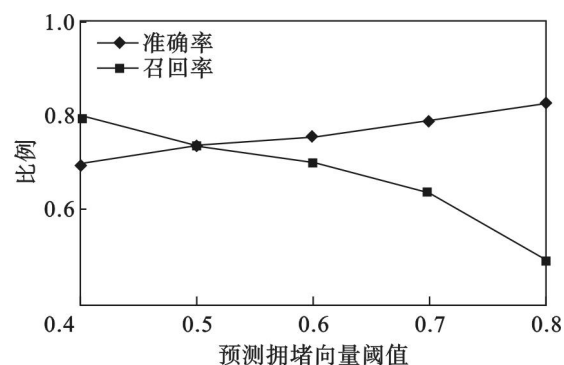


图 3 阈值 β 对准确率和召回率的影响

Fig. 3 The effect of β on precision and recall

大,预测的准确率逐渐增大,而召回率逐渐降低.原因是阈值 β 越大,基于 β 过滤出的路段都是拥堵概率大的路段,得出的拥堵路段越有可能是真实拥堵的路段,因此,准确率越高.而阈值 β 越大,基于 β 过滤出的路段数量越少,包含真实拥堵的路段也就越少,因此,预测的召回率越低.

以上基于真实的 GPS 轨迹数据集,对本文所提出的拥堵路段预测方法的效果进行了验证.从实验中可以看出,随着时间框架长度的增加,预测的准确率呈下降趋势,因此,在实际应用中,设置的时间框架长度不宜过大,以期达到较好的预测效果.同时,应根据实际情况设置合适的阈值 β ,以获得较高的预测准确率和召回率.

4 结 语

本文面向 GPS 轨迹数据,摒弃传统的基于交通流预测和拥堵识别的方法,提出一种基于拥堵向量和拥堵转移矩阵的拥堵路段预测方法.针对路段拥堵具有的时间周期性和时空相关性,对出租车 GPS 轨迹数据进行挖掘和训练,通过建立拥堵向量和拥堵转移矩阵,实现对拥堵路段的预测.基于真实数据集的实验表明,本文所提方法相比于已有的拥堵路段预测方法,预测的准确率有较大提高,并具有较好的扩展性.

参考文献:

- [1] Lécué F, Tucker R, Bicer V, et al. Predicting severity of road traffic congestion using semantic web technologies [M]. Berlin: Springer International Publishing, 2014: 611 – 627.
- [2] Kwoczek S, di Martino S, Nejd W. Predicting and visualizing traffic congestion in the presence of planned special events [J]. *Journal of Visual Languages & Computing*, 2014, 25(6): 973 – 980.
- [3] Schwenke C, Wagner T, Kabitzsch K. Event based identification and prediction of congestions in manufacturing plants [M]. Berlin: Springer Berlin Heidelberg, 2013: 376 – 390.
- [4] Pescaru D. Urban traffic congestion prediction based on routes information [C]// IEEE International Symposium on Applied Computational Intelligence and Informatics. New York, 2013: 121 – 126.
- [5] 姜桂艳, 江龙晖, 王江峰. 城市快速路交通拥挤识别方法 [J]. 交通运输工程学报, 2006, 6(3): 87 – 91.
(Jiang Gui-yan, Gang Long-hui, Wang Jiang-feng. Traffic congestion identification method of urban expressway [J]. *Journal of Traffic and Transportation Engineering*, 2006, 6(3): 87 – 91.)
- [6] Chun-Hsin W, Jan-Ming H, Der-Tsai L. Travel-time prediction with support vector regression [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2004, 5(4): 276 – 281.
- [7] Sun S L, Zhang C S, Yu G Q. A Bayesian network approach to traffic flow forecasting [J]. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 2006, 7(1): 124 – 132.
- [8] Abdulhai B, Porwal H, Recker W. Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks [C]// California Partners for Advanced Transit and Highways (PATH). Berkeley, 1999: 1 – 45.
- [9] Castro P S, Zhang D, Li S. Urban traffic modelling and prediction using large scale taxi GPS traces [C]// Proceedings of the 10th International Conference on Pervasive Computing. Newcastle: Springer-Verlag, 2012: 57 – 72.
- [10] Chen Z, Shen H T, Zhou X. Discovering popular routes from trajectories [C]// 2011 IEEE 27th International Conference on Data Engineering (ICDE). Hannover, 2011: 900 – 911.