基于环境标记约束的不确定轨迹频繁路径挖掘*

张晓滨，张海基
（西安工程大学 计算机科学学院，西安 710048）

摘　要：针对环境约束的不确定轨迹数据的频繁路径问题，设计了一种适应于严格时间约束条件下基于环境约束的位置不确定的移动概率序列挖掘算法（UETFP-PrefixSpan）。算法通过设置类标号把不同环境下的不确定轨迹数据区分开，利用概率支持度对频繁项集进行了重新定义，通过减少某些特定序列模式生成过程的扫描，来减少投影数据库的规模及扫描投影数据库的时间，提高算法效率。测试实验结果表明，改进后的 UETFP-PrefixSpan 算法挖掘结果更符合现实情况，算法执行效率更高。

关键词：序列挖掘；频繁轨迹模式；环境约束；不确定轨迹数据
中图分类号：TP312　　**文献标志码**：A　　**文章编号**：1001-3695(2018)09-2648-03
doi:10.3969/j.issn.1001-3695.2018.09.020

Frequent path mining based on uncertain trajectory of environmental label constraints

Zhang Xiaobin, Zhang Haiji
(*School of Computer Science, Xi'an Polytechnic University, Xi'an* 710048, *China*)

Abstract: Focusing on the frequent path problem about environmental features and uncertain trajectory data, this paper devised a novel algorithm named UETFP-PrefixSpan to mine frequent moving trajectory pattern from environmental features and uncertain data with strict time interval constraints. By setting the class label to distinguish uncertain trajectory data acquisition under different environmental conditions, it redefined the frequent itemsets by using probability support. This algorithm reduced the scale of projected databases and the time of scanning projected databases through reducing scanning of certain specific sequential patterns production. In this way, algorithm efficiency could be raised up. The test results show that the improved UETFP-PrefixSpan algorithm has more realistic on mining result and better efficiency.

Key words: sequence mining; frequent trajectory patterns; environmental constraints; uncertain trajectory data

0　引言

卫星定位技术、无线通信技术和地理信息技术的迅速发展使定位设备在车载以及移动终端上得到广泛使用，不仅用户可以方便地获取个人位置信息，而且信号接收设备可以从定位终端上采集到大量移动对象的轨迹数据。由于一个人的行为方式，爱好在短时间内是不会改变的，所以这些轨迹数据中隐藏着用户在客观世界的活动规律。近年来，大量研究者关注于基于 GPS 的轨迹数据分析研究并取得了很好的效果。

在实际研究过程中，移动对象上传的位置信息受到测量设备精度、噪声干扰、网络传输、空间存储等因素的限制往往带有一定的不确定性[1]，并且不同环境下移动对象的运动模式也有可能不同。因而，现有的面向确定性数据集的模式挖掘方法难以适用到具有环境特征的不确定移动轨迹数据中。

1　相关工作

在频繁模式挖掘方面，文献[2]提出了 Prefix-Span 算法，采用基于投影的分治法来减小数据，以便在挖掘大型序列数据库时具有很高的效率。不过 Prefix-Span 算法只考虑了序列中项目的前后顺序并没有考虑严格的时间间隔。文献[3]证明了在数据量很大的情况下，可以用期望支持度来代替频繁概率，作为挖掘频繁项集的标准，既能减少运算量，又能保证算法的精度。Lin 等人[4]提出了两种用于不确定数据挖掘高效用频繁项集的算法。文献中 PHUI-UP 算法采用产生—检测框架，分层搜索具有高存在概率的高效用频繁项集。并且为了避免多次扫描数据库，文献同时提出了 PHUI-List 算法，该算法采用表结构和集合枚举树直接挖掘频繁项集，避免了候选项集的产生。具体到移动轨迹数据挖掘方面，Geng 等人[5]针对满足最大宽度、最小长度和最小频率的轨迹数据库提出了一种基于深度优先的搜索算法来对轨迹数据进行挖掘。文献[6]引用射频标签阵列来挖掘活动轨迹的频繁轨迹模式。Qiao 等人[7]利用一个标记为空的根节点和一系列原子路段组成的前缀子树重新构建了一个频繁轨迹模式树来对轨迹数据进行频繁模式挖掘。由于种种原因，本文获取的 GPS 轨迹数据并不是确定的，对于不确定轨迹数据的处理，文献[8]考虑到不确定轨迹数据的自身特征，通过引入模糊集方法，提出了一种不确定轨迹数据模式挖掘算法。文献[9]将原始不确定轨迹数据通过基于网格分割面积的不确定轨迹近邻网格概率匹配方法转换为以网格单元表示的概率序列数据。Li 等人[10]针对不确定序列的模式挖掘，利用动态规划的方法来计算模式频度的概率，挖掘出具有间隙约束的概率频繁的时空序列模式。文献[11]基于可能世界语义提出了衡量模式频度。该文献根据许多实际的应用程序产生的不确定性序列数据，建立了两种不确定性序列数据模型，并且制定了符合上述数据的基于概率的频繁序列模式挖掘算法。

收稿日期：2017-04-18；**修回日期**：2017-06-06　　**基金项目**：陕西省教育厅科学研究计划资助项目（14JK1307）

作者简介：张晓滨（1970-），男，副教授，主要研究方向为数据挖掘、个性化服务技术与应用（xiaobinzhangcn@126.com）；张海基（1991-），男，硕士，主要研究方向为数据挖掘、个性化服务技术与应用.

在对具有复杂环境的移动轨迹处理方面,文献[12]采用了带有环境信息值的虚拟参考点代替实际的轨迹点对移动对象进行轨迹预测。文献[13]认为利用语义信息能够更好地描述轨迹数据的真实环境,如轨迹数据所在的场景、轨迹速度和在某个轨迹点的运行模式等。文献[14]提出了一种在噪声环境和高维特征环境因素下具有很好的鲁棒性的最近邻算法。

研究人员虽然利用时空轨迹数据提出了许多关于频繁轨迹模式的算法,但是这些算法仍存在以下两点不足:a)用户场景的环境数据对挖掘结果影响考虑不足;b)现有的轨迹模式挖掘算法对确定的轨迹数据挖掘取得了良好的效果,但对于具有概率属性的轨迹数据的研究内容相对较少。

针对上述问题,本文基于经典序列挖掘算法 Prefix-Span,利用环境的不同把轨迹数据进行了分类标记,同时对频繁项集进行了重新定义,并且针对带时间间隔约束的具有环境特征的不确定轨迹数据模式挖掘问题进行研究,提出了一种适应于不确定轨迹数据的改进算法 UETFP-PrefixSpan。

2　问题描述

本文所研究的轨迹是带有时间和环境属性的位置不确定轨迹流,每条轨迹由多个地点串联而成,每两个地点之间还带有时间差,并且每条轨迹序列都有一个它们所处的环境的类标号。

定义 1　一个轨迹点 p_i 是由一个三元组 (x_i, y_i, t_i) 构成,其中 x_i 和 y_i 是经度和纬度,t_i 是时间戳。

定义 2　定义具有环境特征的不确定移动轨迹形式为 $T = \langle C_i, \{p_0, p_1, \cdots, p_n\}\rangle$,$p_i = \langle x_i, y_i, t_i\rangle$,其中 C_i 为当时所处的环境的类别,$\langle x_i, y_i\rangle$ 为空间位置点的经纬度,$t_i (i = 0, \cdots, n)$ 是相应时间戳,$\forall 0 \leqslant i < n, t_i < t_{i+1}$。通过分组的方式把原始轨迹数据分类,分别挖掘不同环境下的频繁轨迹。

目前主要有四种模型来对不确定移动轨迹数据进行建模和表示,分别是 cylinder [15]、beads [16]、grid 和evolving[17] 模型。本文采用的是 grid 模型。Grid 模型将空间划分为若干不相交的网格,将移动轨迹表示为一组网格序列。

定义 3　本文对具有环境约束的不确定移动时空轨迹的处理是在文献[9]上进行扩展改进,原始不确定轨迹点 (x_i, y_i) 通过转换之后变化成 (U_l, mv_{il})。其中 (x_i, y_i) 为原始不确定轨迹位置数据,U_l 为所匹配的网格单元,mv_{il} 为位置点 (x_i, y_i) 隶属于网格单元 U_l 的隶属度。

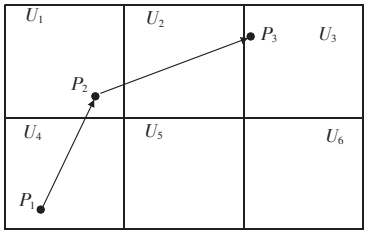

图1　移动轨迹位置点的网格转换图示

如图 1 中的轨迹根据本文的定义转换之后就变成了 $T = \langle C, \{(((U_4, 1)), ((U_1, 0.6), (U_2, 0.15), (U_4, 0.15), (U_5, 0.1)), ((U_2, 0.45), (U_3, 0.55)))\}\rangle$,其中 C 是指获取轨迹序列时的环境因素,比如天气情况。项目集中的数字列是位置点 P_1、P_2、P_3 隶属于网络单元 U_l 的隶属度。

定义 4　在传统的频繁项集挖掘中,项目集 X 的支持计数被定义为包含 X 的事务数目。对于一个不确定的数据集,这样的支持计数很显然不太合适,因为本文不知道项目集 X 是否一定存在。

在本文的数据模型里,一个不确定轨迹数据集 D 由 d 条序列组成,分别是 $t_1, \cdots, t_d$。一条序列 t_i 包含一组项目集。每个项目 x 在 t_i 中都有一个不为零的概率 $P_{t_i}(x)$,表示项目 x 在项目集 i 中存在的可能性。

如果一个项目 x 当且仅当其预期的支持不少于 $\lambda = P_s \times d$(其中 P_s 是一个用户指定的概率支持度阈值)时,x 才被称为一个频繁项。

定义 5　项目 x 的预期支持度 $Se(x)$ 的计算由下面公式给出:

$$Se(x) = \sum_{i=1}^{|D|} \max_{x \in X} P_{t_i}(x) \tag{1}$$

3　改进的 UETFP-PrefixSpan 算法

对轨迹数据的挖掘,现有的算法没有把环境情况考虑在内,但很显然环境的不同可能会导致本文对路径有不同的选择,这对本文的挖掘结果也会造成很大的影响,所以在轨迹数据中加入环境条件是很有必要的。除此之外,已有的序列模式挖掘算法处理的项都是确定性数据并且大多数算法只考虑模式中项目出现的先后顺序,而对项目的出现没有考虑严格的时间间隔约束。所以对于具有环境和时间约束的不确定轨迹数据的挖掘问题,必须对已有的序列模式挖掘算法中的频繁项从概率方面进行重新定义,同时也要进行环境和时间约束方面的改进。

本文针对带环境和时间间隔约束的不确定轨迹数据模式挖掘问题进行研究,提出了适应于不确定轨迹数据的改进算法 UETFP-PrefixSpan。该算法设置了数组 $ft[\,]$ 来存储不同环境下的概率轨迹数据。同时设计了一个信息列表来记录不同轨迹模式的类标号、序列号和相应的时间戳信息。

算法　UETFP-PrefixSpan$(\alpha, l, S|\alpha)$

输入:带环境特征的不确定轨迹数据库 TD,概率相关的最小支持度阈值 λ。

输出:移动轨迹频繁路径集合 $FP = \cup_{k=1}^{K} fp_k$。

参数:α_l 表示长度为 l 的轨迹模式 α,$S|_\alpha$ 为轨迹模式 α 的投影数据库。

a)把输入的带环境特征的不确定轨迹数据根据类标号分别存储到 $ft[\,]$ 数组中;

b)分别扫描分组后的投影数据库 $S|_\alpha$;

(a)如果 $l = 0$,扫描投影数据库 $S|_\alpha$,根据文中频繁项的定义,探测 $S|_\alpha$ 中的频繁网格单元集合 $Fg = \{fg_m, 1 \leqslant m \leqslant M\}$;所探测到的频繁网格单元 fg_m 即为长度等于 1 的频繁轨迹模式 α_1,同时记录轨迹模式 α_1 的信息表 α_1_list =[类标号,序列号 ID,时间戳];

(b)如果 $l > 0$,扫描投影数据库 $S|_\alpha$,令 Δt 为一条轨迹上相邻两个点的时间差,根据时间间隔产生 α 的后缀序列 β,探测 β 中的网格单元集合 $\beta' = \{\beta'_m, 1 \leqslant m \leqslant M\}$,记录网格单元 β'_m 的信息表 β'_m_list;计算每一个候选项 β' 的预期支持度 $Se(\beta')$,若满足最小支持度条件 $Se(\beta') \geqslant \lambda$,就把频繁项 β' 连接到 α 形成频繁序列 α',α' 为长度等于 $(l+1)$ 的频繁模式 α_{l+1},更新相应的信息表 α_{l+1}_list;

c)对于新生成的序列模式 α_{l+1},构建相应的投影数据库 $S|_{\alpha+1}$,调用 UETFP-PrefixSpan $(\alpha, l, S|_\alpha)$。

4　实验与分析

4.1　数据集与环境

本文通过文献[18]中的 GSTD 时空轨迹数据合成算法,

产生了轨迹规模为 3 000 条,平均长度为 20 ~ 40 个轨迹点的移动轨迹测试数据集。3 000 条数据中被随机分成了两部分,分别为 1 000 条和 2 000 条用来模拟雨雪和晴朗两种环境情况。将该算法与 UTFP-PrefixSpan 算法进行了对比。

4.2 测试结果和分析

本文从以下四个方面进行了对比:

a)对比如图 2 所示。由图 2 可以看出,UTFP-PrefixSpan 随着最小支持度的不断增大,算法的运行时间逐渐减少。UETFP-PrefixSpan 随着给定的概率支持度阈值 P_s 的增大,算法的运行时间也逐渐减小。在 P_s 为 0.55 时,可以发现算法的运行时间有一个明显的减少,随后的时间变化就比较平缓。这表示当 P_s 值为 0.55 时挖掘出来的结果是比较符合实际的。P_s 持续增大对挖掘结果的影响也不是太大。并且从图中可以看出 UETFP-PrefixSpan 算法要比 UTFP-PrefixSpan 算法的运行时间短。

b)对比如图 3 所示。由图 3 可以看出,随着支持度的逐渐增大,利用 UTFP-PrefixSpan 算法挖掘结果数量不断减少。UETFP-PrefixSpan 随着给定的概率支持度阈值 P_s 的增大,挖掘结果也逐渐减少,并且挖掘结果比 UTFP-PrefixSpan 算法更少,这是因为在挖掘的过程中 UETFP-PrefixSpan 算法利用概率对频繁项进行了重新定义,减少了某些不符合现实应用的项,使挖掘结果更有意义。

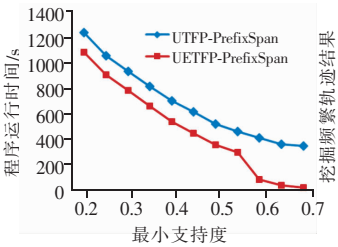

图2 不同支持度下运行时间比较

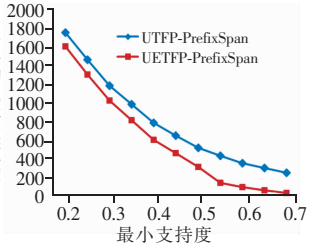

图3 不同支持度下频繁轨迹挖掘结果比较

c)对比将 UTFP-PrefixSpan 算法中的最小支持度设置成 0.55,将 UETFP 算法中的概率支持度阈值也设置成 0.55。从图 4 中可以发现,两种算法随着轨迹条数的递增运行时间也逐渐增加,但可以很明显地看出 UETFP-PrefixSpan 算法的增长速度更加缓慢。

d)对比如图 5 所示。规定参与测试的数据为 3 000 条,将 UTFP-PrefixSpan 算法中的最小支持度设置成 0.55,将 UETFP-PrefixSpan 算法中的概率支持度阈值也设置成 0.55。随着平均轨迹长度的不断增加,两种算法的运行时间均不断增加。并且 UETFP-PrefixSpan 算法的运行时间要比 UTFP-PrefixSpan 更短。

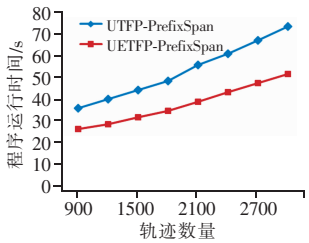

图4 不同轨迹数量下运行时间比较

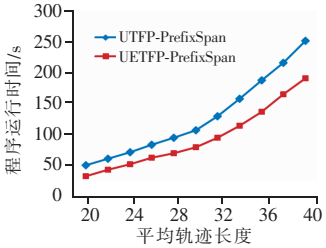

图5 不同平均轨迹长度下运行时间比较

5 结束语

本文提出了一种基于环境约束的不确定移动轨迹数据的频繁路径的挖掘算法,其基本过程是首先对原始轨迹根据多层分类规则把原始轨迹分类,然后再运用改良的 UETFP-PrefixSpan进行严格时间间隔约束条件下的频繁移动轨迹模式挖掘。该算法考虑了现实场景中的环境因素,在挖掘部分加入了环境特征,使挖掘出的频繁轨迹更符合实际,更能描述出人的行为模式。

参考文献:

[1] 李佳佳,王波涛,王国仁,等. 不确定移动对象的查询处理技术研究综述[J]. 计算机科学与探索,2013,7(12):1057-1072.

[2] Pei Jian, Han Jiawei, Mortazavi-Asl B, *et al*. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth[C]// Proc of the 17th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2001: 215-224.

[3] Wang Liang, Cheung W L, Cheng R, *et al*. Efficient mining of frequent item sets on large uncertain databases[J]. IEEE Trans on Knowledge & Data Engineering, 2012, 24(12): 2170-2183.

[4] Lin J C W, Gan Wensheng, Fournier-Viger P, *et al*. Efficient algorithms for mining high-utility itemsets in uncertain databases[J]. Knowledge-based Systems, 2016, 96(C): 171-187.

[5] Geng Xiaoliang, Arimura H, Uno T. Pattern mining from trajectory GPS data[C]//Proc of IIAI International Conference on Advanced Applied Informatics. Washington DC: IEEE Computer Society, 2012: 60-65.

[6] Liu Yunhao, Zhao Yiyang, Chen Lei, *et al*. Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays[J]. IEEE Trans on Parallel and Distributed Systems, 2012, 23(11): 2138-2149.

[7] Qiao Shaojie, Han Nan, Zhu W, *et al*. TraPlan: an effective three-in-one trajectory-prediction model in transportation networks[J]. IEEE Trans on Intelligent Transportation Systems, 2015, 16(3): 1188-1198.

[8] 李帆,夏士熊,张磊. 基于模糊理论的不确定轨迹模式挖掘[J]. 微电子学与计算机,2011,28(8):70-73.

[9] 王亮,胡琨元,库涛,等. 位置不确定移动时空轨迹频繁模式挖掘[J]. 小型微型计算机系统,2014,35(12):2659-2663.

[10] Li Yuxuan, Bailey J, Kulik L, *et al*. Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases[C]//Proc of International Conference on Data Mining. Washington DC: IEEE Computer Society, 2014: 448-457.

[11] Zhou Zhao, Da Yan, Ng W. Mining probabilistically frequent sequential patterns in uncertain databases[J]. IEEE Trans on Knowledge & Data Engineering, 2014, 26(5): 1171-1184.

[12] 夏卓群,胡珍珍,罗君鹏. EAVTP:一种环境自适应车辆轨迹预测方法[J]. 小型微型计算机系统,2016,37(10):2375-2379.

[13] 廖律超,蒋新华,邹复民,等. 一种支持轨迹大数据潜在语义相关性挖掘的谱聚类方法[J]. 电子学报,2015,43(5):956-964.

[14] Wiebe N, Kapoor A, Svore K. Quantum nearest-neighbor algorithms for machine learning[J]. Quantum Information & Computation, 2014, 15(3): 318-358.

[15] Xiao Jianchuan, Xu Li, Lin Limei, *et al*. A privacy-preserving approach based on graph partition for uncertain trajectory publishing [C]//Proc of the 15th International Symposium on Parallel and Distributed Computing. 2016: 285-290.

[16] Kuijpers B, Othman W. Trajectory databases: data models, uncertainty and complete query languages[J]. Journal of Computer and System Sciences, 2010, 76(7): 538-560.

[17] Jeung H, Lu H, Sathe S, *et al*. Managing evolving uncertainty in trajectory databases[J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(7): 1692-1705.

[18] Theodoridis Y, Nascimento M A. Generating spatio-temporal datasets on the WWW[J]. ACM SIGMOD Record, 2000, 29(3): 39-43.