文章编号:1004-9037(2014)05-0840-09

移动对象无冗余周期模式发现算法

许寿全 皮德常

(南京航空航天大学计算机科学与技术学院,南京,210016)

摘要:技术的发展进步和需求的多样性,产生了大量的数据,数据背后隐含模式的发现对问题的深入研究起到关键性的作用。为了挖掘移动对象的周期运动模式,提出了一种无冗余周期模式(Non-redundant period patterns, NRPP)发现算法。为解决噪声因素的影响,在现有方法基础上,引入相似性因子,删减冗余周期模式,使挖掘出的模式精简、准确。算法为不同模式设置不同阈值,不但解决了稀有项问题和组合爆炸问题,而且也使得挖掘效率更高效。采用公开移动对象数据实验,结果表明算法能高效挖掘出移动对象的周期模式。

关键词:移动对象;周期模式;数据挖掘;冗余模式;稀有项

中图分类号:TP2;TP9 **文献标志码**:A

Algorithm of Non-Redundant Periodic-Frequent Patterns for Moving Objects

Xu Shouquan, *Pi Dechang*

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

Abstract: The development of technology and multiple requirements produce huge data. It is critical to mine patterns behind the data for farther in-depth study. To mine period-frequent patterns of moving objects, an algrithm of non-redundant period patterns(NRPP) is proposed. The algorithm set different limited conditions to solve the combinatorial explosion problem and the rare item problem, and it is more efficient. To prevent noise and other uncertainties, similarity-based pattern matching method is introduced. By facilitating existing methods, the proposed method can be more concise and accurate. The experimental results using open data show that the method can efficiently excavate the periodic-frequent patterns.

Key words: moving objects; period pattern; data mining; redundant patterns; rare item

引 言

随着科学技术的日益发展,人们获取数据的方式越来越多样化,数据量也越来越庞大。特别是GPS、无线、通信、网络科技的成熟和广泛应用,使得收集到大量的移动对象数据,例如:手机移动信息,车辆位置,动物迁徙等。挖掘数据背后的规律、模式对人类科研和生活提供巨大的方便,可以帮助指挥交通、研究动物习性等。

一般来说,对移动对象的挖掘应用可分为3类:(1)聚类[1]。从一堆训练数据集中找出聚类模型,然后用这个模型来预测移动对象所属种类;(2)异常检测[2]。发掘持续正常移动模式中出现异常的移动行为,进而发现异常事件或发现环境的变化;(3)移动对象模式分析,包括周期模式、关联模式、频繁模式等[3]。本文主要研究移动对象周期模式,通俗地讲即固定的地理位置、固定的时间间隔重复性的发生某种活动。例如:秃鹰在每年的10月下旬迁徙到南美,第二年的3月中旬返回阿拉斯加。

通过对历史数据的研究分析,周期行为模式更直观、更精确地反应移动对象历史数据,是对原始数据的一种总结。例如:通过对燕子移动轨迹分

基金项目:国家自然科学基金委员会和中国民用航空局联合(U1433116)资助项目;中央高校基本科研业务费(NZ2013306)资助项目。

收稿日期:2013-07-10;**修订日期**:2013-12-23

析，发现该物种白天出去觅食晚上回巢，冬天飞到南方，夏天飞到北方。可以说该物种具有以天为单位的周期行为和以年为单位的周期行为[4]。因此可以用这些模式替代原始数据来节省存储空间，同时这些模式还可以为移动轨迹预测提供辅助依据。虽然挖掘出的周期模式压缩了原始数据，但是周期模式仍然存在冗余。例如表 1，模式 P_1 已经包含 P_2 中的信息，所以 P_2 模式存在冗余性。

表 1　周期模式

Table 1　Periodic pattern

模式	周期	*ids*
$P_1\{A,B\}$	2	1,3,5,8,10
$P_2\{A,B\}$	2,4	1,5

为了解决冗余性问题，已经有人提出基于三元关系理论[5]和最小生成子[6]概念的周期代表算法，同时保存了所有信息[7]，但存在稀有项问题和组合爆炸问题以及不是很好的抗干扰能力。本文提出一种改进的无冗余周期模式发现算法(Non-redundant period patterns，NRPP)，动态获取支持度阈值，引入相似因子，弥补当前研究上的不足。

1　相关定义

对于给定的按时间顺序的数据集，经过预处理可以将其按时间段组合，如图 1 所示，其中左图为时间(单位为 s)对应出现项，右边每隔 0.1 ms 作为一个时间点得到的项集，预处理后为时间点对应项集：

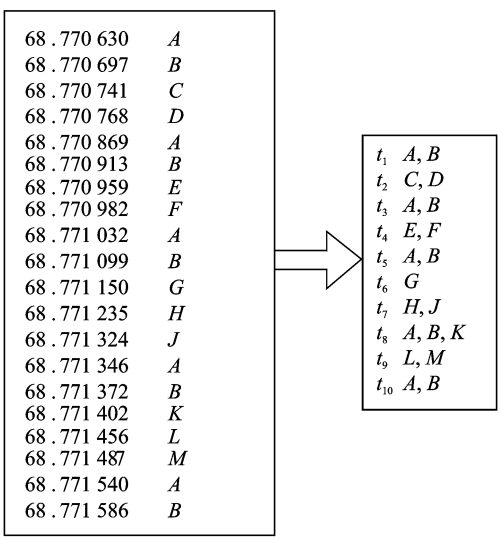

图 1　以 0.1 ms 分段的项集

Fig. 1　Itemsets per 0.1 ms

1.1　频繁周期模式

定义 1　假设项目集 $I=\{i_1,i_2,\cdots,i_n\}$，对于 $X=\{i_j,\cdots,i_k\}\subseteq i,j<k$，且 $j,k\in[1,n]$，那么 X 称为模式或项目集。对于事务 $w=(tid,X)$，其中 tid 代表事务 id，即事务的唯一标识符，X 代表模式。关于 I 的事务数据库 $W=\{t_1,t_2,\cdots,t_m\}$，$m=|W|$，其中 $|W|$ 是 W 的事务总数。

定义 2　给定模式 X 和周期 p，模式 X 关于 p 的周期模式可以表示为 $\mathrm{Cycle}(t_o,p,l,X)$，$t_o$ 代表起始位置，l 代表模式出现次数。$\mathrm{Cycle}(t_o,p,l,X)$可表示为下式：$\mathrm{Cycle}(t_o,p,l,X)=\{t_k\in D|X\subseteq t_k,k=o+p*i,0\leqslant i<l,X\not\subset t_{o-p},X\not\subset t_{o+p*l}\}$，其中 D 为事务数据库，$0\leqslant o<n,2\leqslant l\leqslant n,n=|D|$。

定义 3　假设模式 X 及 Cycle，$C=\{(o_1,l_1),\cdots,(o_k,l_k)\}$，如果所有周期为 p，并且是延续的 $(\forall(o_i,l_i),(o_j,l_j)\in C,o_i<o_j)$不出现重叠$(\forall(o_i,l_i),(o_j,l_j)\in C,o_i+(p*(l_i-1))<o_j)$，那么称其为周期模式 $P(X,p,s,C)$

定义 4　假设最小支持度阈值 minsup，只有当周期模式 P 的支持度大于该阈值，模式 P 才是频繁周期模式。例如，只有当 $s\geqslant$minsup，$P(X,p,s,C)$ 才是频繁[8]的。

假设用图 1 中的数据，设置最小支持度阈值为 2，那么得到的频繁周期模式如图 2 所示。

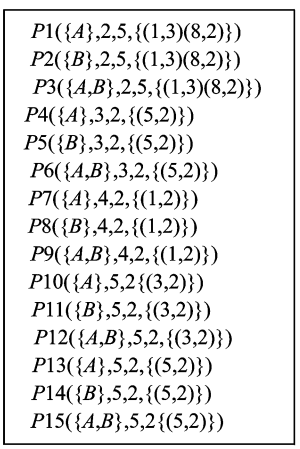

图 2　频繁周期模式集

Fig. 2　Frequent periodic patterns

图 2 中出现了大量的冗余，频繁项的组合连同频繁项被看做频繁周期模式，例如 $P1,P2,P3$ 出现在相同的事务中，$P3$ 包含了 $P1,P2$ 中所有的信息，因此 $P1,P2$ 为冗余模式。另外，短周期模式包含长周期模式，例如 $P3$ 周期为 2 包含 $P9$ 周期为 4 的模式，$P9$ 为冗余模式。

1.2 稀有项问题和相似度

所谓稀有项[9]就是出现比较少的项。某些模式出现频率高且周期短,而那些出现频率低且周期长的模式称为稀有项,稀有项可能更有意义。通过设置支持度阈值和周期距离阈值容易挖掘出频繁周期项,但稀有项往往被忽略。稀有项一般具有较低的支持度和较大的周期距离,设置唯一的支持度阈值和周期距离阈值很难挖掘出这些稀有项。如果设置较低的支持度阈值和较高的周期距离阈值,会引发组合爆炸问题[10]。同时,一些无意义的模式,因设置的支持度阈值过低、周期距离阈值过高而被误认为频繁周期模式,降低了频繁周期模式的可信度。

为了增加算法健壮性,最大程度降低噪声的干扰,不同于轨迹相似性度量[11],对模式 X 提出相似度参数 $Pr(X)$[12]:假设 $IP^X=\{P_k,\cdots,P_l\}\subseteq P^X=\{P_1,P_2,\cdots,P_n\}$,其中 $k\leqslant l$, $\forall P_i^X\in P^X$, $Per(P_i^X)\leqslant$maxprd 且 $s(P_i^X)\leqslant$minsup,其中 Per 为模式周期长度,maxprd 为最大周期距离阈值,那么模式 X 的相似度为 $Pr(X)=\dfrac{IP^X}{P^X}$。如果相似度大于用户设置的相似度阈值,那么该模式为冗余模式,否则不是。对于模式 X, $Pr(X)\in[0,1]$,若 $Pr(X)=0$,则两个模式没有相同项;若 $Pr(X)=1$,则两个模式出现时间点完全相同,将保存为代表模式。对于 $0<Pr(X)<1$ 的模式即使部分时间点不一样,但如果大于阈值仍视为冗余周期模式,做删除处理。

2 无冗余频繁周期模式发现算法

算法大致可分为 3 个模块:(1)获取周期项算法 GetPeriods(遍历整个数据库将大于阈值的项作为周期项提取出来);(2)挖掘三元关系(包括项集,周期集,事务 ids);(3)获得无冗余周期模式(某些三元关系被另一个三元关系包含的做删除处理)。

2.1 周期项获取

对于原始数据,首先要进行预处理,如图 1 中所示,得到时间-项目集数据库。然后遍历得到的数据库,记录每个项出现的时间,如果出现个数大于指定的支持度阈值则作为周期项保留。在设置支持度阈值时,对于项 i 采用动态设置的方法,即:$S_p>S\times\gamma$,其中 S_p 代表周期为 p 的项 i 的出现次数,S 为项 i 出现的总次数,γ 为百分比。该方法改善了设置固定值作为周期阈值的缺点,即如果设置过大,出现的稀有项会被删除掉,如果设置过小,导致组合爆炸。该方法很好地解决了稀有项问题和组合爆炸问题。

图 2 中的数据集可以用三元关系表示,如 $P1$ ($\{A\}$,2,5,$\{(1,3),(8,2)\}$)可以表示为$(A,2,t_1)$, $(A,2,t_3)$, $(A,2,t_5)$, $(A,2,t_8)$, $(A,2,t_{10})$。如果设置支持度阈值为 2,那么图 2 可以用表 2~5 表示。

表 2 $p=2$ 的三元关系表

Table 2 Triadic table under under conditional $p=2$

i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
A	1		1		1			1		1
B	1		1		1			1		1

表 3 $p=3$ 的三元关系表

Table 3 Triadic table under conditional $p=3$

i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
A					1			1		
B					1			1		

表 4 $p=4$ 的三元关系表

Table 4 Triadic table under conditional $p=4$

i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
A	1					1				
B	1					1				

表 5 $p=5$ 的三元关系表

Table 5 Triadic table under conditional $p=5$

i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
A			1		1			1		1
B			1		1			1		1

为了得到表 2~5 的数据,采用 GetPeriods 算法对原始数据作处理,GetPeriods 算法流程图如 3 所示。

其中 GetTriples 算法为获取项 i 周期为 p 的事务 ids,GetTriples 算法流程图为如图 4。其中 TS 格式为(i,p,ids),i 为项,p 为周期,ids 为项 i 周期为 p 所出现的事务。Len 为项 i 周期 p 出现的时间点的个数,Next 为下一个项目出现的时间点。

以上两个算法将原始数据库提取出形如表 2~5 中表示法,即(i,p,ids)。

2.2 最大三元关系挖掘

由算法 1 得到的为单个项对应某个周期出现

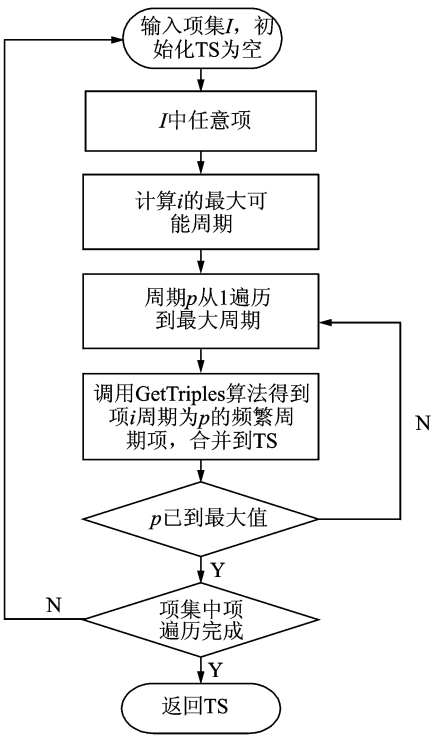

图 3　GetPeriods 算法

Fig. 3　Flowchart of GetPeriods algorithm

的 ids，这样的模式存在大量冗余信息并且很难发现多个项之间的关系。如图 2 中的 A 和 B 具有完全相同的模式特点，应该将其合并。为了使挖掘出的信息更加简洁，引入三元关系[5]，它能保存最大三元关系，并且包含所有的信息。

根据文献[13]中的方法，得到最大三元关系数据库 TC，结果如表 6 所示。数据库中不存在更大的三元关系中的项，能够包含图 2 中三元关系中的项。例如数据库中不包含$\{A,B,C\}$与图 2 中的模式$\{A,B\}$具有相同的周期和 ids。该关系库包含所有原始信息，并且比原始信息更加简洁、易懂。

表 6　最大三元关系

Table 6　Max-triadic

模式	周期长度	时间点
P_1	$\{2\}$	$\{t_1,t_3,t_5,t_8,t_{10}\}$)
P_2	$\{2,4\}$	$\{t_1,t_5\}$
P_3	$\{2,5\}$	$\{t_3,t_5,t_8,t_{10}\}$
P_4	$\{2,3,5\}$	$\{t_5,t_8\}$
P_5	$\{2,3,4,5\}$	$\{t_5\}$

得到的最大三元关系库可以转换为周期模式，形如 $P(X,p,s,C)$。例如：$P_2(\{A,B\},\{2,4\},\{t_1,t_5\})$只包含周期为 4 的模式，转换为($\{A,B\}$,4,2,

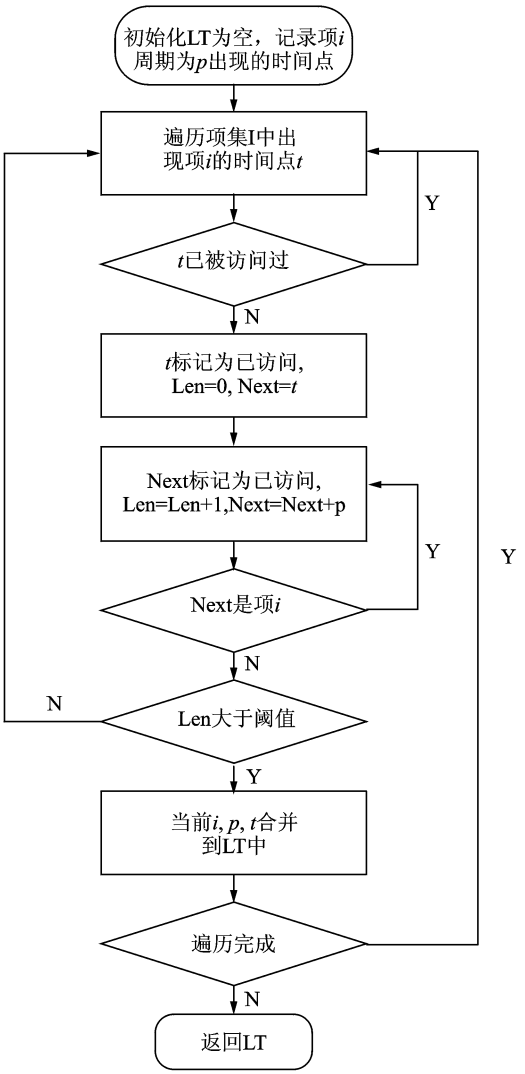

图 4　GetTriples 算法流程图

Fig. 4　Flowchart of GetTriples algorithm

{(1,2)})(Period＝4，Support＝2，起点为 1，长度为 2)，对应图 2 中 P_9，不包含多余信息 P_7，P_8，降低了冗余性。

虽然相对图 2 简洁很多，但是仍然存在冗余信息。例如：P_1 和 P_2 具有相同的模式，P_2 出现的时间点包含在 P_1 出现时间点内，所以 P_1 包含 P_2 所有信息，P_2 为冗余项。

2.3　无冗余周期模式的获得

一般来说，周期小的模式往往包含周期大的模式，时间节点数多的模式包含时间节点数少的模式。例如周期为 2 的模式 $ids\{t_1,t_3,t_5,t_8,t_{10}\}$包含周期为 4 的模式 $ids\{t_1,t_5\}$。假设三元关系(I,P,T)，如果不存在(I',P',T')使得 $I=I'$，$P'\subset P$，$T\subset T'$，则称(I,P,T)为最小周期三元关系。表 6

$P_2(\{A,B\},\{2,4\},\{t_1,t_5\})$模式与 $P_1(\{A,B\},\{2\},\{t_1,t_3,t_5,t_8,t_{10}\})$模式都为$\{A,B\}$，因为$\{2\}\subset\{2,4\},\{t_1,t_5\}\subset\{t_1,t_3,t_5,t_8,t_{10}\}$，所以 P_1 为最小周期三元关系，而 P_2 被 P_1 包含不是最小三元关系，做删除处理。

现实中的数据往往存在噪音的干扰，例如周期较大的三元关系出现的事务标识集合 id 大部分都包含在周期小的三元关系 id 中。由于噪音的干扰，导致部分 id 不被包含。本文引入相似度的概念，如果包含在内的 id 占总数的比例大于某个值 η，将该三元关系视作冗余关系做删除处理。采用提出的方法对表 3 进行删减，结果如表 7 所示。

表 7 无冗余周期模式

Table 7 Periodic pattern of non-redundance

模式	周期长度	时间点
$M\{A,B\}$	$\{2\}$	$\{t_1,t_3,t_5,t_8,t_{10}\}$

表 7 相对于图 2 和表 6 更加简洁、清晰，将其转换为周期模式 Cycle 形式为 $P(\{A,B\},2,5,\{(1,3),(8,2)\})$对应图 2 中的 $P3$，可见它包含了图 2 中所有其他周期项。图 5 为 GetMinPer 算法流程图。

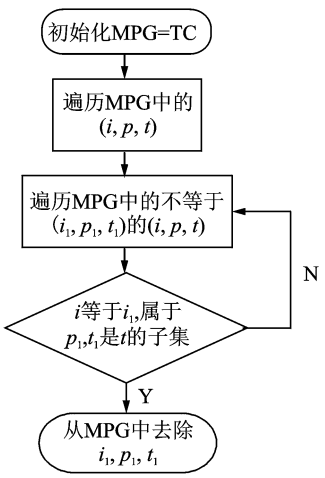

图 5 GetMinPer 算法流程图

Fig. 5 Flowchart of GetMinPer algorithm

输出的 MPG 即为所求频繁周期模式，并且不包含冗余信息。

3 实验分析

实验采用的公开数据集 1 为 Starkey 项目中获取的动物迁徙数据集中的经纬度数据，并将其转换成 UTMGridEast（参考网站）和 UTMGridNorth（参考网站），即动物迁徙轨迹（www. fs. fed. us/pnw/ starkey/data/tables/），该数据集包含物种 C,D,E 的迁徙轨迹，其中 C = cattle，D = deer，E = elk。不同的物种迁徙规律不同，所以将其分开研究，分成 3 个数据集存储在 C. TXT，D. TXT 和 E. TXT，分别包含 49 669，70 582 和 166 885 个点。由于原始数据集合中包含的位置属性比较多，本次实验只需动物的位置坐标信息，即经纬度信息。

3.1 物种 catte 实验分析

首先对物种 C 的移动数据做预处理，设置每 300 个时间测点作为一个时间段，同一时间段内的位置点具有相同的时间测点。并且每个时间段内如果出现相同的位置点，只保留一次。

为物种 C 动态设置最小支持度占出现总次数比例 $\gamma=0.03$，连续周期数比例设置 3%，总周期数占总出现次数比例设置为 10%，包含在最小周期三元关系中的比例阈值 η 设为 0.6，如图 6 所示，横纵坐标分别代表 UTMGridEast 和 UTMGridNorth。作为对比分别设置固定支持度 2 和 7，其他参数不变，如图 7，8 所示。

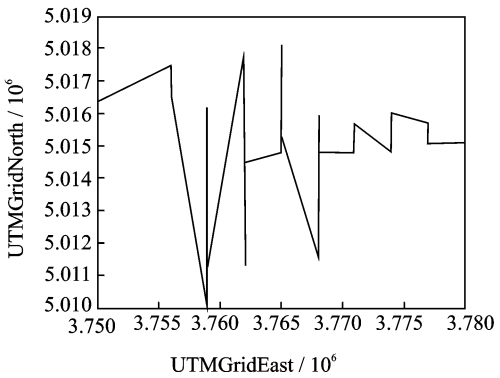

图 6 动态设置阈值

Fig. 6 Variational threshold

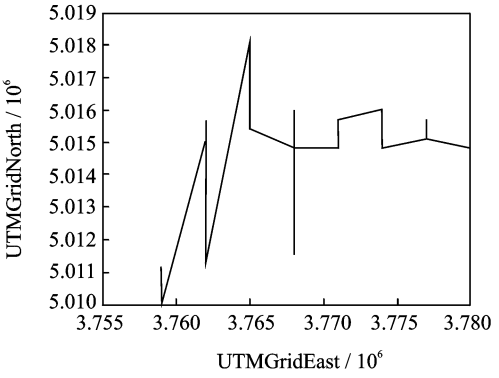

图 7 固定阈值 7

Fig. 7 Fixed threshold 7

图 7 设置阈值过高，将某些重要模式遗漏掉，如(375 000，5 016 300)，影响对移动对象研究的准确性。图 8 设置阈值过低，导致挖掘出的模式比较多，对不必要周期模式过多研究造成资源的浪费，并且有可能影响对移动对象周期行为的判断。

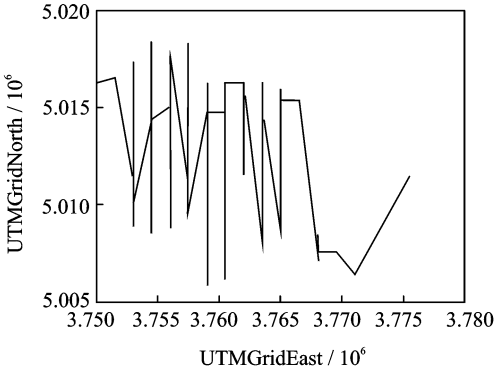

图 8　固定阈值 2

Fig. 8　Fixed threshold 2

相似性因子 η 的引入使得周期行为相似的模式被缩减，从而减少相似模式的存在，降低冗余性。设置不同的 η，控制着相似程度，例如设 $\eta=1$，得到频繁周期模式如表 8 所示。

表 8　$\eta=1$ 物种 C 的频繁周期模式

Table 8　Frequent patterns of C under conditional $\eta=1$

坐标	周期	时间点
(376 500，5 017 500) (377 400，5 016 000)	250	184，684，934，434
(376 500，5 017 500) (377 700，5 015 100)	250　240	202
(3 76 500，5 017 500)	250　240	202，198
(376 500，5 017 500) (375 900，5 016 300)	250　240	198
(376 500，5 017 500) (375 900，5 016 300) (377 400，5 016 000)	240	918，678，438，198

相似性因子 $\eta=0.6$ 得到频繁周期模式如表 9 所示。

表 9　$\eta=0.6$ 物种 C 的频繁周期模式

Table 9　Frequent patterns of C under conditional $\eta=0.6$

坐标	周期	时间点
(376 500，5 017 500) (377 400，5 016 000)	250	184，684，934，434
(376 500，5 017 500) (375 900，5 016 300) (377 400，5 016 000)	240	918，678，438，198

设置不同的 η 得到的频繁周期模式数目也不同，$\eta=1$ 得到的频繁周期模式数为 1 980，而 $\eta=0.6$ 得到频繁周期模式数 252，大大降低了模式的冗余性。

3.2　物种 deer 实验分析

对物种 D 的移动数据做预处理，设置每 300 个时间测点作为一个时间段，同一时间段内的位置点具有相同的时间测点，并且每个时间段内如果出现相同的位置点，只保留一次。

为物种 D 动态设置最小支持度占出现总次数比例 $\gamma=0.1$，连续周期数比例设置 5%，总周期数设置为 8%，包含在最小周期三元关系中的比例阈值 η 设为 0.6，动态设置时，如图 9 所示横纵坐标代表 UTMGridEast 和 UTMGridNorth。作为对比，分别设置固定支持度 10 和 50，其他参数不变，如图 9，10 所示。

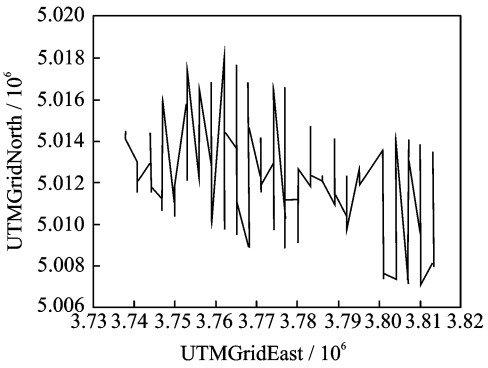

图 9 动态设置阈值

Fig. 9　Bariational threshold

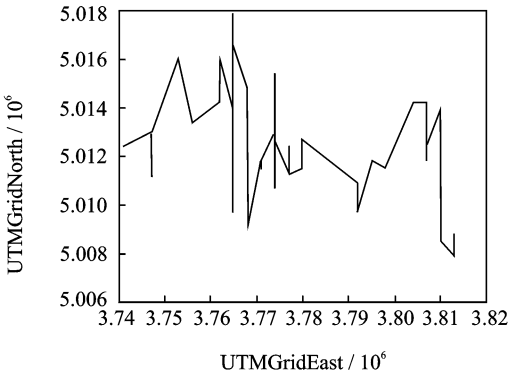

图 10　固定阈值 50

Fig. 10　Fixed threshold 50

图 10 阈值设置过高，遗漏很多稀有项，如位置点(373 800，5 014 500)，虽然挖掘出大致趋势，但是会遗漏重要稀有项。图 8 设置阈值过低，产生组

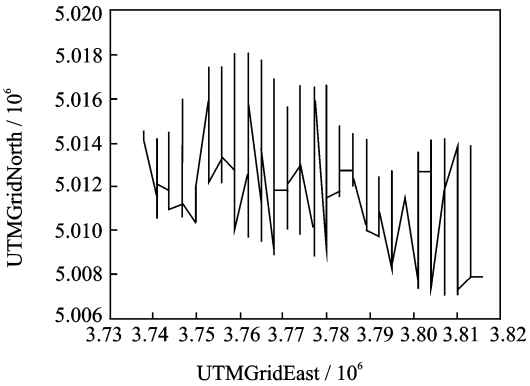

图 11 固定阈值 10

Fig. 11 Fixed threshold 10

合爆炸,将多余位置点当成频繁周期点挖掘出来,如位置点(374 100,5 014 200),这为后期处理带来很多冗余。

为了删除冗余周期模式,分别为 η 设置 1 和 0.6,$\eta=1$ 得到的频繁周期模式如表 10 所示。

表 10 $\eta=1$ 物种 D 的频繁周期模式

Table 10 Frequent patterns of D under conditional $\eta=1$

坐标	周期	时间点
(380 400,5 007 300)	380	634, 1 015, 635, 254, 255, 1014
(380 400,5 007 300)	380, 400, 410	254, 255
(377 100,5 012 100)	380	1 385, 624, 1004, 1 005, 1 384, 625
(377 100,5 012 100)	380,350	1 384
(377 100,5 012 100)	380,340	625,624
(377 700,5 013 300)	540	9, 13, 549, 553, 1 089, 1093

设置 $\eta=0.6$ 得到频繁周期模式如表 11 所示。

表 11 $\eta=0.6$ 物种 D 的频繁周期模式

Table 11 Frequent patterns of D under conditional $\eta=0.6$

坐标	周期	时间点
(380 400,5 007 300)	380	634, 1 015, 635, 254, 255, 1014
(377 100,5 012 100)	380	1 385, 624, 1 004, 1 005, 1 384, 625
(377 700,5 013 300)	540	9, 13, 549, 553, 1 089, 1 093

可见,$\eta=0.6$ 的频繁周期模式中不存在 $\eta=1$ 中的冗余周期模式。在 $\eta=0.6$ 时,数据库记录数为 355,$\eta=1$ 时的记录数为 272。可证明相似度的引入可以删除部分冗余周期模式。

3.3 物种 elk 实验分析

对物种 E 的移动数据做预处理,设置每 300 个时间测点作为一个时间段,同一时间段内的位置点具有相同的时间测点。并且每个时间段内,如果出现相同的位置点,只保留一次。

为物种 E 动态设置 $\gamma=0.05$,连续周期比例设置为 5%,总周期数设为 8%,包含在最小周期三元关系中的比例阈值 η 设为 0.6,如图 12 所示,作为对比,分别设置固定支持度 20 和 6,其他参数不变,如图 13,14 所示。

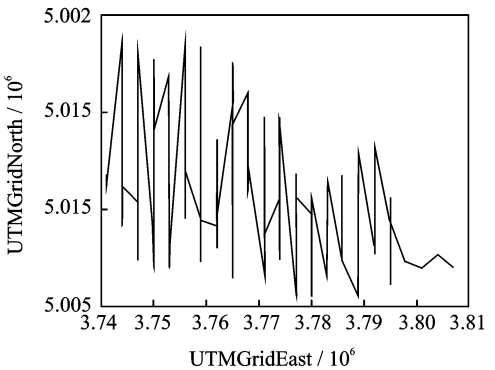

图 12 动态设置阈值

Fig. 12 Variational threshold

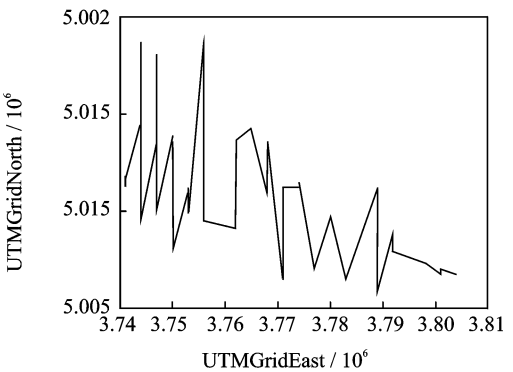

图 13 固定阈值 20

Fig. 13 Fixed threshold 20

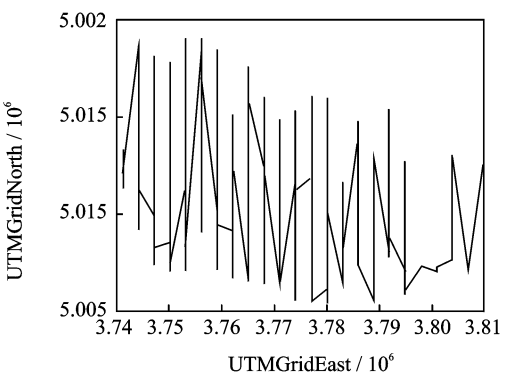

图 14 固定阈值 6

Fig. 14 Fixed threshold 6

图 13 设置固定阈值过大，挖掘出的周期漏掉了稀有项，图 14 设置阈值过小，不仅运行缓慢浪费时间，而且挖掘出很多没意义的周期模式，影响对移动轨迹的研究。

物种 E 设置不同 η，设置 $\eta=1$ 得到的频繁周期模式如表 12 所示。

表 12　$\eta=1$ 物种 E 的频繁周期模式

Table 12　Frequent patterns of E under conditional $\eta=1$

坐标	周期	时间点
(379 200,5 008 800)	310，290，280	2 006
(378 000,5 010 300)	310	244，1 490，554，864，560，1180，1 484，870，1 174，2 110，2 420，1 800，1 794
(378 000,5 010 300)	310,320	870
(376 500,5 010 300)	310,320	3 097，2 132，2 127，2 123，1 817，3 096，2 442

设置 $\eta=0.6$ 得到频繁周期模式如表 13 所示。

表 13　$\eta=0.6$ 物种 E 的频繁周期模式

Table 13　Frequent patterns of E under conditional $\eta=0.6$

坐标	周期	时间点
(378 000,5 010 300)	310	244，1 490，554，864，560，1 180，1 484，870，1 174，2 110，2 420，1 800，1 794

可见 $\eta=0.6$ 中只保留了 $\eta=1$ 中的一个周期模式，其余模式都作为冗余模式删除，再次证明本文方法可以减少挖掘出的周期模式冗余。

4　结束语

在获取数据手段日趋多样化的今天，移动对象周期模式挖掘的研究正如火如荼地进行。但是挖掘出的模式存在很大的冗余性，并且挖掘方法不够灵活，解决噪音干扰能力有限。通过实验分别验证了动态设置阈值的方法能够减少挖掘出模式的冗余，解决稀有项问题和组合爆炸问题，引入相似度的方法能够抗击一定的噪音干扰，使得相似周期模式被删减，进一步降低周期模式的冗余性。由此得出，提出的移动对象 NRPP 发现算法能够降低挖掘出的周期模式冗余性，并且解决了稀有项问题和组合爆炸问题，提高了算法的健壮性。

参考文献：

[1] 刘金勇，郑恩辉，陆慧娟. 基于聚类和微粒群优化的基因选择方法[J]. 数据采集与处理，2014，29(1)：83-89.
Liu Jinyong，Zheng Enhui，Lu Huijuan. Gene selection based on clustering method and particle swarm optimization[J]. Journal of Data Acquisition and Processing，2014，29(1)：83-89.

[2] 郭迎春，吴鹏，袁浩杰. 基于自投影和灰度检索的视频帧中异常行为检测[J]. 数据采集与处理，2012，27(5)：612-619.
Guo Yingchun，Wu Peng，Yuan Haojie. The anomalous behavior detection in video sequence based on self-casting histogram and gray histogram[J]. Journal of Data Acquisition and Processing，2012，27(5)：612-619.

[3] Li Zhenhui，Han Jiawei，Ji Ming，et al. Movemine：Mining moving object data for discovery of animal movement patterns[J]. ACM Transactions on Intelligent Systems and Technology，2011，2(4)：1-37.

[4] Li Zhenhui，Ding Bolin，Han Jiawei，et al. Mining periodic behaviors for moving objects[C]//KDD′10 Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，USA：ACM，2010：1099-1108.

[5] Rudolf W. The basic theorem of triadic concept analysis[J]. SpringerLink，1995，12(2)：149-158.

[6] Nicolas P，Yves B，Rafik T，et al. Efficient mining of association rules using closed itemset lattices[J]. Information Systems，1999，24(1)：25-46.

[7] Patricia L C，Aurelie B，Alexander T，et al. Debugging embedded multimedia application traces through periodic pattern mining[C]//EMSOFT′12 Proceedings of the Tenth ACM International Conference on Embedded Software. New York，USA：ACM，2012：13-22.

[8] 汤春明，王培义. 在线挖掘数据流闭合频繁项集 CMNL-SW 算法的研究[J]. 数据采集与处理，2012，27(4)：508-513.
Tang Chunming，Wang Peiyi. CMNL-SW algorithm study on online mining closed frequent itemsets over data stream[J]. Journal of Data Acquisition and Processing，2012，27(4)：508-513.

[9] 丁艳辉，王洪国，高明，等，一种发现有价值的稀有数

据关联规则的算法[J]. 山东师范大学学报:自然科学版,2005,20(4):17-19.
Ding Yanhui, Wang Hongguo, Gao Ming, et al. Finding significant rare data association rule algorithm[J]. Journal of Shandong Normal University: Natural Science, 2005,20(4):17-19.

[10] 刘禹,朱智源,关强,等. 基于试验设计的 RFID 应用组合测试优化研究[J]. 自动化学报,2010,36(12):1674-1680.
Liu Yu, Zhu Zhiyuan, Guan Qiang, et al. Research on experimental-design-based RFID application combinatorial testing optimization[J]. Acta Automatica Sinica,2010,36(12):1674-1680.

[11] 赵洪斌,韩启龙,潘海为. 移动对象轨迹时空相似性度量方法[J]. 计算机工程与应用,2010,46(29):9-12.
Zhao Hongbin, Han Qilong, Pan haiwei. Spatio-temporal similarity measure for trajectories on road networks[J]. Computer Engineering and Applications,2010,46(29):9-12.

[12] Uday T K, Krishna P R. An alternative inerestingness measure for mining periodic-frequent patterns [J]. SpringerLink, 2011, 6587: 183-192.

[13] Loic C, Jeremy B, Celine R, et al. Closed patterns meet n-ary relations [J]. ACM Transactions on Knowledge Discovery from Data,2009, 3(1): 1-36.

作者简介:许寿全(1988-),男,硕士研究生,研究方向:数据挖掘,E-mail: xushouquan120@126.com;皮德常(1971-),男,教授,博士生导师,研究方向:数据挖掘、大数据处理,E-mail: nuaacs@126.com。