基于周期表的时空关联规则挖掘方法与实验

柴思跃[1,2]，苏奋振[1]，周成虎[1]
（1. 中国科学院地理科学与资源研究所 资源与信息系统国家重点实验室，北京 100101；
2. 中国科学院研究生院，北京 100049）

摘要：地理现象的周期性往往掩盖了许多地学规律，这也是地学数据挖掘的一个主要内容。本文以周期表设计了一种时空层次关联规则挖掘方法——PRules-Miner。模型利用周期表的表现形式对时空数据进行组织，并通过两步挖掘过程发现具有“遥相关”地理事物间的变化模式。模型算法分为 3 个步骤：(1)过滤周期表内无序数据：逐行地提取多周期内时空状态的频繁项，生成新的时空频繁状态表；(2)基于向下闭合引理，对时空频繁状态表中的对象进行时空拓扑匹配，得到时空关联规则候选集；(3)对于候选数据集进行时空拓扑验证，得到时空关联规则集。为证明模型算法的可靠性，应用 PO.DAAC 提供的 20 年 AVHRR Product 016 海表面温度遥感反演数据集和国家气象科学院提供的南京地区降水逐日数据资料，研究大洋暖池与南京降水间的时空关联规则。实践表明，这种挖掘方法具有以下特点：(1)算法基于面向对象思想，对地理对象状态进行独立描述。因此，所得时空关联规则与时空粒度无关，并能够挖掘出时空粒度不一致的地物间的关联关系。(2)算法使用笛卡尔积得到在时空拓扑阈值内匹配的时空候选集，并可以发现时域、空域均不邻接的事物间的时空关联规则，即时延不确定的地理现象的相互关联。

关键词：数据挖掘；关联规则；时空数据；层次挖掘；周期表
DOI：10.3724/SP.J.1047.2011.00455

1 引言

时空数据是地理现象在时空维度上的采样，包含时间、空间、属性 3 个基本特征，是地理信息系统(GIS)的基本成分。由于缺少强有力的数据处理工具理解历史数据，近年来遥感和 GIS 技术的广泛使用所积累的大量时空数据并不能被现有 GIS 系统有效利用，因此，数据挖掘技术被引入到地理学领域，旨在从海量地学数据中发现其中隐含的模式与知识。关联规则挖掘作为数据挖掘技术的重要组成部分，从地理数据向真实世界反馈 IF-THEN 形式的知识。而时空关联规则挖掘，是关联规则的子领域，能够寻找包含了时间-空间拓扑关系的关联规则。规则形如“生活在加拿大西北部的驯鹿群大部分时间不迁徙，仅在一定区域内活动”[1]，规则也可形如“若台湾东北部海域的盐度从 0.15psu 上升至 0.25psu，则下个月台湾东北部海域的海水温度将从 0℃ 上升至 1.2℃”[2]。时空关联规则的表述是多样的，其复杂性由现实世界中已知的或所需的知识所决定。

时空关联规则中，时间是首先被考虑的问题。这里引用 Li 等文中的实例来证明时间对于规则的重要性。“存在煎蛋≥咖啡（支持度＝3%，置信度＝80%）的规则，这说明 3%的消费者同时购买煎蛋和咖啡，并且买了煎蛋的消费者中有 80%的人同时购买了咖啡。然而，在现实数据中，“煎蛋≥咖啡”消费模式大部分出现在早上 7 点至上午 11 点的时间段内。在这段时间内，规则具有 40%以上的支持度。然而在其他时间段，该规则的支持度小于 0.005%”[3]。可以看出，时间范围的界定有助于提高规则被发现的概率。同理，空间范围在规则挖掘中同样具有限定功能。

然而，现有关联规则挖掘方法大都将时间数据与空间数据割裂开进行处理，这些方法或只能挖掘序列规则或只局限于利用事物空间关系挖掘静态空间关联，对于时空关联规则却无能为力。目前，对于统一时间与空间维度的联合数据挖掘方法的研究尚不成熟[4-7]。

另一方面，在地理现象中，存在具有远距离相互关联的模式。尤其是大气科学中，“遥相关”

收稿日期：2010-11-16；**修回日期**：2011-06-07.
作者简介：柴思跃(1985-)，男，北京人，硕士，研究方向：时空数据挖掘。E-mail：chaisy@lreis.ac.cn

(Teleconnection)被定义为描述不同地区间大气异常环流的相关的概念(反相位或同相位),是大气环流的重要模态之一[8]。例如,厄尔尼诺与南方涛动之间存在典型的遥相关关系。因此,广义角度上,"遥相关"是存在一定距离的地物间可能存在的系统性相关关系[9]。

目前,在数据挖掘领域,对于这类具有"遥相关"关系的关联规则挖掘方法研究尚处于起步阶段。但面对时空数据的日益增长,我们迫切地需要一种方法来发现这种事物间的关联关系,为更深入地研究地理现象,以及地球系统过程提供时空关联规则指引。本文设计了一种基于周期表的时空关联挖掘模式,并且应用 PO.DAAC 提供的 1981 年末至 2001 年初共 20 年的海表面温度遥感反演数据集和同时间内的南京地区降水数据逐日资料为数据基础,挖掘大洋暖池与南京降水间的时空关联规则。

2 时空关联规则挖掘的相关研究与问题

2.1 相关研究

频繁模式与关联规则挖掘的命题最先于 1994 年由 Agrawal 提出[10]。经过十几年发展,多种关联规则挖掘方法被设计出来,目前,挖掘算法可以分为 4 大类:(1)根据频繁模式与关联规则挖掘根据数据类型的不同可分为布尔型与数量型;(2)根据规则涉及的数据维数的不同分为单维与多维;(3)根据规则集所涉及的抽象层次的不同分为单层关联规则与多层关联规则;(4)根据模式与规则间的相互关系分为完全、最大、闭合型[11-12]。从关联规则的意义出发,大量研究致力于从数据库中挖掘具有时间或空间拓扑关系的关联规则[13-18]。因此,关联规则也可以分为 3 大类:时间关联规则挖掘、空间关联规则挖掘、时空关联规则挖掘。

时间关联规则挖掘是关联规则挖掘中的重要组成,用于挖掘包含时间信息或序列信息的知识。例如,"购买了佳能数码照相机的顾客很可能在一个月内购买 HP 彩色打印机。"时间关联规则包含挖掘序列模式、相同时间序列集和时间规则 3 类[2]。这些方法在对于时间的表示上,可以分为时间戳与时间段两种方式。经典的序列关联规则挖掘算法是 SPADE[19]和 GSP[20]。目前,时间关联规则挖掘,大都考虑局限于利用 Allen 提出的时间拓扑关系[21]作为挖掘约束条件。时间规则挖掘算法中,Ozden 等利用事物的周期重复特点设计的周期表挖掘方法可以显著提升规则支持度,从而发现更隐蔽的规则[3,22-23],例如,规则"购买火鸡的顾客会同时购买南瓜饼"会出现于 11 月 80%的天数中。

空间关联规则挖掘研究在于发现空间实体间的相互作用关系,例如"90%靠近海滩的房子价格都高"[24]。其中,空间实体间的相互作用、空间依存、因果或共生等模式。空间关联规则挖掘利用 GIS 空间分析方法对数据进行空间维度的拓扑计算,在特定邻域内搜索与发现空间依赖模式[25-28]。

对于同时挖掘数据的时空信息,还没有成熟的算法。Mennis 等利用序列挖掘应用在带有时刻和空间描述的数据库中寻找关联规则[29]。Lee 对具有时空特性的移动对象的路径轨迹进行挖掘,获取移动对象的时间模式,其实质就是带有地理坐标的序列挖掘[2]。Tao 等简略表述了时空关联规则,提出了用"如果事物在 R_i,t 时间内,而后在 $t+\tau$ 时间,事物出现于 R_j"表示时空关联,记作$(r_i,\ \tau,\ p)\geqslant r_j$,并应用贪婪算法和 FM-PCSA 技术加速[30]。Verhein 和 Chawla 基于移动对象时空数据库对移动对象在空间区域之间的移动时空关联规则进行挖掘,探索其间的关联模式,完善了 Lee *et al* 的思想,将时空关联规则表达形式完善为 STAR:$A(r_i,\ TI_s,\ p)\geqslant B(r_j,TI_e,q)$[1]。Verhein 的方法用源-通路-汇集地三种状态描述空间格网,主要针对群体目标事物的移动规律,但存在对于非连通路径和异常斑块无法解释的问题。Huang 等学者以 RPPI 挖掘方法为基础,提取了台湾附近海域相邻时间点内不同空间上的海盐与温度关联规律,得到时空过程规则:$A(r_i,\ t,\ p)\Rightarrow B(r_j,t+1,q)$[2]。但该实验将空间描述设定为以台湾岛为中心,将空间描述固化为海面温度的一个属性,既没有体现出空间变化,同时时间属性的限定性也有限。

2.2 存在的问题

首先,我们回顾 Agrawal 对关联规则的定义[31]。设存在项的集合 $I=\{I_1,I_2,\ldots,I_m\}$。对于相关数据库中存储的各个行的集合 T,存在 $T\subseteq I$。其中每个事务 T 都有一个唯一标识符 TID。存在事务集 $A\subseteq T$,$B\subseteq T$,并且 $A\cap B=\Phi$,满足给定支持度阈值 min_sup 与置信度阈值 min_conf 后,A、B

同时发生的概率 $P(A\cup B)\geqslant$min_sup,$P(B|A)\geqslant$min_conf。称 $A\Rightarrow B$ 为一条关联规则。

关联规则已成熟应用于购物篮分析中。然而,地理知识纷繁复杂,地理数据的类型也并不统一。这种复杂性造成时空关联规则挖掘与购物篮分析具有显著差异性。这首先表现在时空关联规则中包含的时空拓扑关系决定了时空关联规则具有独特的表达特点;其次,时空数据库中数据结构的异质性,造成时空数据类型既包含一般数据库表结构数据,同时又包含影像数据;再次,与商业关联规则相比,时空关联规则的生成与验证都要更加复杂。可以说,时空关联规则挖掘的难点在于以下 4 个方面:

(1)如何表达复杂、多样的时空关联规则。

(2)如何在数据操纵过程中,保持时空一致性。

(3)对于海量数据事物在时间-空间中的多种形态,如何发现其中存在的频发状态。

(4)如何匹配事物发生的时空顺序,并验证时空关联规则的拓扑关系,使其具有合理的地学意义。

3　基于周期表的时空关联规则挖掘方法与实验分析

3.1　周期表与时空状态的描述

地理现象总是存在周而复始的特点。在事物运动周期中,总是表现出相同或相似的变化规律。即在多个周期中,地理事物在特定的时间段上会表现出相同或相似的状态,而这些状态集则是地理规则的基础。例如,在强南极涛动(AAO)年,中国长江流域夏季降水偏多,中国北方的沙尘天气和太平洋台风发生频数较少[32-34]。事例的前半句结论说明,在 AAO 年份长江流域夏季降水多于多年夏季降水平均值的现象是频繁出现。因此,利用地理现象的周期性特点,我们可以从时空状态入手,对时空关联规则加以描述。

定义 1:时空状态,也就是地理事件,是指特定时间在特定地点的地理事物的属性状态。

定义 2:周期,是指事物运动完成一次重复所需的时间。

定义 3:时间分辨率是指在同一区域进行的相邻两次采样的最小时间间隔。

在现有的地理信息数据库中,地理事件按照采样时间按顺序记录为二维表结构。每条地理事件存放着描述时间、空间、属性信息,且每条记录存在一个唯一标识。但由于顺序记录数据,造成对数据时间信息解读异常困难(图 1)。在图 1 中,我们以大洋暖池平均温度与面积大小两种属性为例,可以发现两种属性的震荡规律并不明显。因此,顺序记录的方式并不能完全表现数据中存在的时空知识。

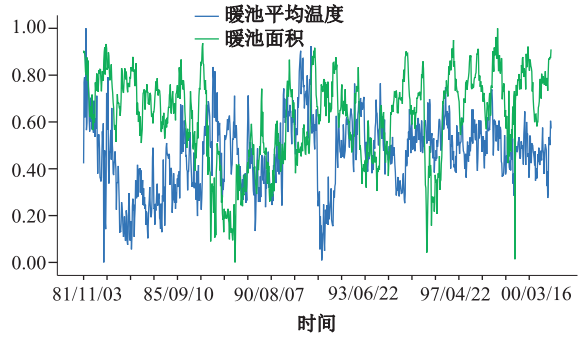

图 1　事物属性序列均一化示例图

Fig. 1　Example chart of attribute sequential data for some objects

定义 4:设顺序采样集为 $D=\{d_1, d_2,\cdots, d_n\}$,包含周期个数为 n。在相同采样频率时,每个周期内采样点数量为 m。周期表定义为矩阵 P_{m*n},且周期表内元素 $p_{ij}=d_{n*(i-1)+j}$, $i\in[1,n-1]$, $j\in[1,m]$。矩阵 P_{m*n}形如:

因此,我们引入周期表的定义,通过周期表将顺序记录的地理对象集 $T=\{(TI_1,L_1,A_1),(TI_2,L_2,A_2),\cdots,(TI_m,L_m,A_m)\}$转换为周期表集的形式:$T=\{P_{TI},P_L,P_A\}$。这使得原先数据库中以序列方式记录的时间信息、空间信息与属性信息被转化为对象化周期表记录集。这种转化不但使得数据的存储与管理易于管理,而且使得数据在周期中的时间信息具有了固定的二维坐标,实现了数据表现形式与概念形式的统一。

另一方面,空间数据结构的不一致性是数据挖掘前期必须克服的障碍。空间数据库中存储的影像信息需要被统一转化为二维表结构。而这种转化的过程是提取地理对象的时空信息的记录过程,在此仅以图表为例,加以展示(如图 2、表 1)。

周期表结构的行向量存储了各个周期的数据,这些行向量包含了地理对象周期性变化的数据。在特定行内,存在一些经常出现的状态,叫做频繁出现的时空状态。这些状态正是形成时空关联规则的基础(表 2、3)。

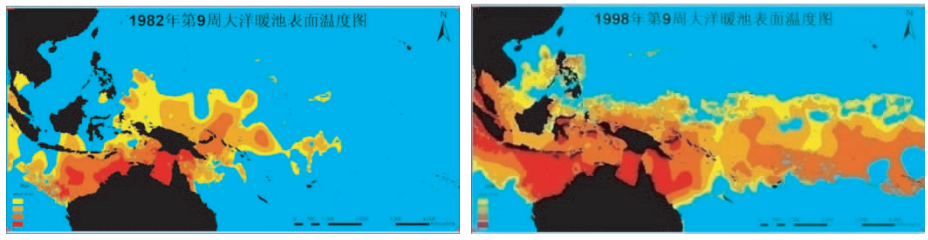

图 2 海表面温度场反演数据的大洋暖池提取示例图

Fig. 2 Example chart of Warm Pool extraction result from sea surface temperature inversion data

表 1 大洋暖池时空信息提取结果

Tab. 1 Spatio-temporal information extraction result from Warm Pool

ID	最高温度	中心经度	中心纬度	平均温度(℃)	面积(km^2)	时间	年	月
1	32.25	156.6195	2.6100	28.96	42550865	1981-11-03	1981	11
2	33.6	158.5687	1.4289	29.45	42101391	1981-11-10	1981	11
3	33.15	160.898	0.1748	29.48	42420670	1981-11-16	1981	11
4	32.7	161.4557	−1.170	29.24	40765241	1981-11-24	1981	11
…	…	…	…	…	…	…	…	…
998	32.1	162.1003	−7.351	29.17	42717593	2001-02-07	2001	2

表 2 大洋暖池数据周期表集——以平均温度距平数据集为例

Tab. 2 Example of period tables: anomaly average temperature of Warm Pool (0.1℃)

	第一周	第二周		第五十二周
1981			…	0.371107
1982	0.42943	0.393908	…	0.261577
1983	0.05837	−0.03516	…	−0.17508
1984	−0.23567	−0.31682	…	−0.41475
1985	−0.468	−0.34879	…	0.185367
…	…	…	…	…
2001	0.04054	0.054448	…	−0.2501

表 3 南京旬降水距平数据表(降水量单位:0.1mm)

Tab. 3 Example of period tables: anomaly average precipitation of Nanjing (0.1mm)

	第一旬	第二旬	…	第三十六旬
1981	−69.4	411.7	…	−49.8
1982	−76.4	−77.7	…	−52.8
1983	121.6	−36.7	…	378.2
1984	579.6	252.3	…	−82.8
1985	−94.4	325.7	…	−113.8
…	…	…	…	…
2001	−98.4	77.7	…	−116.8

针对问题 3、4,为快速搜索并发现时空关联规则,我们引入向下封闭引理(Downward Closure Lemma):若闭集 I 不频繁出现,则必定存在 I 集的任意非空子集是不频繁的[35]。其中,“频繁”一词是指 I 中元素出现的频率大于一定阈值。我们将引理进行推广,使 I 集合中的元素包含时间与空间信息,则存在:(1)时空关联规则中的频繁出现的时空状态必定是频繁的。(2)时空关联规则存在于由时间拓扑约束频繁集与空间拓扑约束频繁集的交集中。这说明,时空关联规则中存在的状态一定是频繁出现的时空状态,也说明时空关联规则可以表示为空间规则规范下的时间规则或时间规则规范下的空间规则。在这里,需要强调的是只有同时满足支持度和置信度的规则才称为强关联规则。

3.2 PRules-miner 挖掘方法设计

通过对时空关联规则的定义,我们设计了 PRules-Miner 挖掘方法用于挖掘时空关联规则,下文将对 PRules-Miner 进行详细介绍。

3.2.1 挖掘策略

PRules-Miner 采用递进式挖掘时空关联的模式,分阶段解决归纳知识,即先寻找时空频繁状态,再对状态集进行连接与验证,得到时空关联规则。应用 Apriori 挖掘算法实现时空频繁状态提取处理。但对于时空关联规则的生成与验证使用匹配技术,以事物发生的时空顺序组织频繁出现的时空状态,以此来反映时空关联模式,并使之具有合理的地学意义。在语义学中,频繁出现的时空状态通

常是描述知识的主体结构,例如“北京夏季降水占年降水量的 72.5%”[36]。但这种表述只能描述单一事物。另一方面,利用时空拓扑关系是时间拓扑与空间拓扑的耦合结果的特点,通过分布挖掘时态规则与空间规则,才能达到挖掘时空关联规则的目的。一般来说, Allen 模型常用于表达时间拓扑关系,而九交模型则用于表达空间关联关系。时空关联规则的种类是时间关联规则与空间关联规则的笛卡尔积。

3.2.2　时空关联规则挖掘流程

PRules-Miner 可分为 3 个主要步骤(图 3):

(1)数据准备

(2)频繁出现的时空状态的挖掘

(3)时空关联规则挖掘。

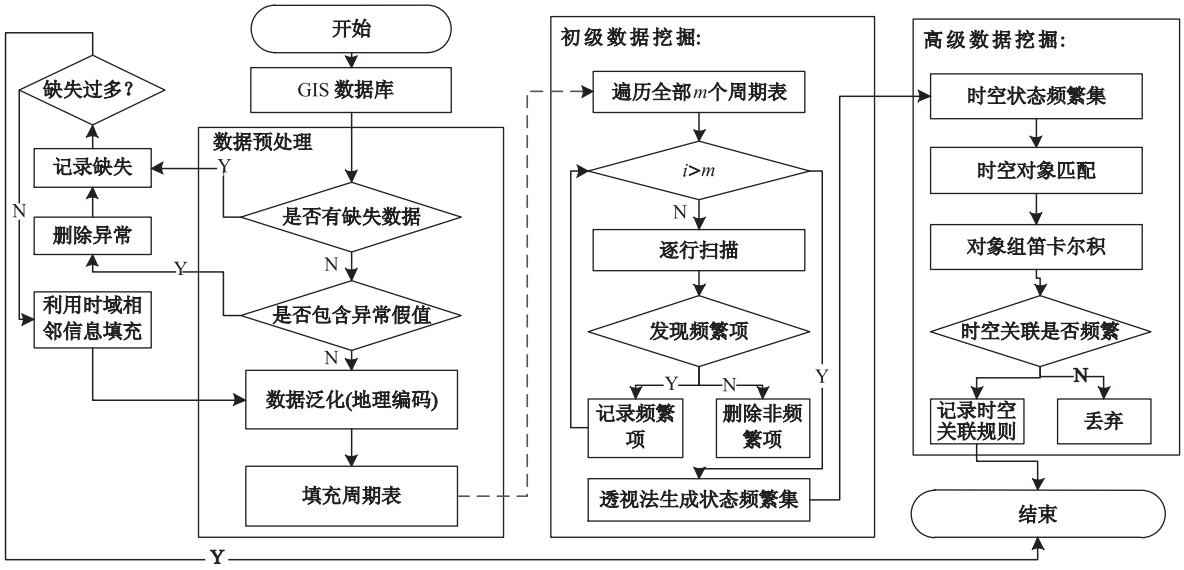

图 3　PRules-Miner 挖掘流程图

Fig. 3　Flow chart of PRules-Miner process

首先,进行数据清洗工作,填充数据中的缺失值,并进行数据的分类与编码。为更好地衡量数据状态,我们采用时间段内多年平均值的距离作为统计值,并以平均值与标准差为分类体系,对各个周期表内数据进行分类。例如,对平均温度距平数据表中第九周数据进行分类,得到阈值表 4。

表 4　平均温度距平分类阈值

Tab. 4　Example of classification threshold: anomaly average temperature of Warm Pool

类别	公式	下限	上限
1	X<Mean-2 * SDV		< −0.336225
2	Mean-2 * SDV≤X<Mean-SDV	≥ −0.336225	< −0.1681125
3	Mean-SDV ≤X<Mean+SDV	≥ −0.1681125	< 0.1681125
4	Mean+SDV≤X< Mean+2 * SDV	≥0.1681125	< 0.336225
5	X≥ Mean+2 * SDV	≥0.336225	

初级数据挖掘在整个数据库内进行。依照历史周期表结构,在表内进行初步挖掘得到各个时间段内,地理对象的单一属性的频繁状态。例如,“1987、1988、1990、1993、1994、1995、1998 年的第九旬中,大洋暖池的分布面积大致为 32 919 279km^2”或“1982、1988、1990、1993、1994、1998 年的第九旬,大洋暖池的平均温度约为 29.14℃”。利用向下闭合引理,只有在一个时段内全部属性都频繁的事物,才记录一条频繁出现的时空状态。经过初级挖掘,汇总频繁的时空状态,得到频繁状态数据集(表 5)。其中,各年份中的数值为各属性频繁状态叠加后结果。简单来讲,即是多年内事物反复出现了相同状态。例如,“1988、1990、1993、1994、1998 年的第九旬中,大洋暖池的分布状态相似。其面积大致为 32 919 279km^2,平均温度约为 29.14℃,且重心位置在本时段多年平均重心的东北部,即位于巴布

亚新几内亚以东海域"。反映在表 5 中,若标注出现这种状态的年份为 T,反之标注为 F。因此,在同一时段内可能出现多种运动状态,但由于本文数据所限,这种多态性在表 5 中并没有表现出来。

表 5 大洋暖池频繁状态集

Tab. 5 Example of frequent status set of Warm Pool

ID	时间段	面积类型	1981	1982	…	2000	2001	重心方位偏移距平	平均温度距平类别	支持度(%)	置信度(%)
24	第二十一周	5	F	F	…	F	F	东南	5	21	73
10	第九周	5	F	T	…	F	F	东北	5	27	81
43	第三十五周	5	F	F	…	F	F	不变	5	27	81
14	第十三周	5	F	T	…	F	F	东北	5	21	73
51	第四十二周	5	F	F	…	F	F	西北	5	23	76
54	第四十四周	5	F	F	…	F	F	东	5	21	73
50	第四十一周	5	F	F	…	F	F	北	4	21	73
63	第五十一周	5	F	F	…	F	F	东北	5	21	73

数据挖掘中的第 3 步,通过给定时空拓扑关系构建时空关联候选集(表 6)。构建方法是将所有满足时空拓扑关系并且频繁出现的时空状态进行笛卡尔积运算,生成时空关联规则候选集。挖掘过程则是在时空候选集中寻找满足给定时空拓扑关系下时空支持度与时空置信度的规则。

表 6 时空关联规则挖掘候选集

Tab. 6 Candidate set of spatio-temporal association rules

暖池ID	暖池时间段	暖池面积类别	暖池1981	暖池1982	…	暖池2001	重心方位偏移距平	平均温度距平类型	降水ID	降水时间段	降水类别	降水1981	…	降水2001
10	第九周	5	F	T	…	F	东北	5	1	第一旬	1	T	…	T
43	第三十五周	5	F	F	…	F	不变	5	1	第一旬	1	T	…	T
10	第九周	5	F	T	…	F	东北	5	2	第二旬	1	F	…	F
43	第三十五周	5	F	F	…	F	不变	5	2	第二旬	1	F	…	F
10	第九周	5	F	T	…	F	东北	5	6	第六旬	1	F	…	F
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
43	第三十五周	5	F	F	…	F	不变	5	42	第三十六旬	1	T	…	T

时空关联规则挖掘是一个交互、复杂、反复的过程,挖掘方法的选取和挖掘效果都会由于问题的不同、数据组织的不同而存在一定的差异。整个挖掘过程是由一系列前后紧密相连的阶段构成,每个阶段的结果均作为下一个阶段的输入。因此,当前阶段结果的好坏直接影响下一个阶段的操作,最终会影响最后的挖掘结果。

3.3 时空关联规则挖掘实验与结果分析

实验选用 PO.DAAC 提供的 NOAA-AVHRR 传感器的海洋表面温度再分析数据。影像大小为 2048×1024 像元,数据周期为 8 天,时空覆盖度为 20 年,数据起始时间为 1981 年 11 月。我们从中提取大洋暖池的时空信息,分别用时空、重心经度、重心纬度、平均温度、覆盖面积描述。对于另一地理对象中国南京地区逐日降水资料,降水量单位为 0.1mm,且经度误差小于 0.1mm。由于两数据在时空尺度上存在差异。因此,依据大气学中旬降水量的概念将降水资料进行转换,尽量减小两套数据间的时空尺度差异。

阈值设定为:状态支持度为 30%,状态置信度为 70%。在实验中时空拓扑关系为 1 年后,时空支持度为 20%,时空置信度为 70%。挖掘结果如表 7。

本文以第一条规则为例,可以解释为:若暖池在第九周(3 月 1 - 7 日)时出现位置比往年偏东北方向,面积比往年显著增大,即面积大于10 913 873 km^2且平均温度高于旬多年平均温度 0.34℃,则次年的第六旬(3 月 1 - 10 日)南京地区降水出现显著小于多年旬平均值,即降水量小于旬多年均值 45.2mm。其支持度为 21%,置信度为 80%。

表 7　时空关联规则挖掘结果

Tab. 7　Results of PRules-Miner

ID	Spatio-Temporal Association Rules	ST_sup	ST_conf
1	(WP,第 9 周,平均温度>0.34℃,东北,面积>10913873km2)　1Year_Later (Najing,第 6 旬,降水距平<−45.2mm) => True	0.21	0.8
2	(WP,第 9 周,平均温度>0.34℃,东北,面积>10913873km2)　1Year_Later (Najing,第 1 旬,降水距平<−23.3mm) => True	0.24	0.8
3	(WP,第 9 周,平均温度>0.34℃,东北,面积>10913873km2)　1Year_Later (Najing,第 2 旬,降水距平<−40mm) => True	0.24	0.8
4	(WP,第 9 周,平均温度>0.34℃,东北,面积>10913873km2)　1Year_Later (Najing,第 7 旬,降水距平>−38.2mm 且降水距平<−184mm) => True	0.24	0.8
5	(WP,第 35 周,平均温度>0.55℃,东北,面积>21067383km2)　1Year_Later (Najing,第 19 旬,降水距平<−128.5mm) => True	0.21	0.8
6	(WP,第 35 周,平均温度>0.55℃,东北,面积>21067383km2)　1Year_Later (Najing,第 30 旬,降水距平<−61.1mm) => True	0.21	0.8

对比关联规则 1 中所对应的数据,由于位置关系是以非数值型数据表示方位(图 5),因此,不将其纳入数值型数据分析中。通过距平均值较远的特征出现的概率大于所定义的 30%,面积、平均温度也出现此特征的条件概率大于 80%,即对两组数据在方差的时空分布上出现同步特征。通过观察图 5 可以发现 1982、1988、1990、1991、1993、1994、1998 年中,两曲线均处于高点。同时,在 1990、1991、1993、1994、1998 年的第九周时,暖池中心偏离多年平均中心位置,且都处于多年平均中心的东北方向。由于数据量有限,其余年份的点上并没有表现出明显的同步特征。结合大洋暖池状态图,我们可以发现在 1988、1990、1993、1994、1998 年大洋暖池的分布状态相似。其面积大致为 32 919 279km^2,平均温度约为 29.14℃,较 20 年年均面积 28 707 909km^2 高 14.67%,较 20 年平均温度 28.99℃高 0.5%。这五年中心位置均处于(162.16°E,5.84057°S 左右,位于 20 年平均中心位置(160.7477°E,7.23134°S)东北部的巴布亚新几内亚以东海域。其时空状态见图 4。

在暖池状态基础上加入了南京地区第 6 旬降水均一化统计。面积与平均温度处在相关,而第 6 旬降水,与暖池数据出现很好的反向规律,即暖池数据为峰值时,降水数据出现波谷。依据先验知识给定的大洋暖池与降水间的时空拓扑关系,如果将第 6 旬降水量向右移动一个刻度,则暖池与降水间会出现相似的趋势。而时空关联规则 1,恰恰说明了在暖池出现位置比往年偏东北方向,面积处于峰值且平均温度也处于峰值时,则次年的第 6 旬(3.1−3.10)降水出现谷值。其支持度为 21%,置信度为 80%。数据表明,大洋暖池分别在 1990、1993、1994、1998 年 4 年中出现,比往年偏东北方向,面积比往年显著增大且平均温度比往年显著升高,且在第二年即 1991,1994,1995,1999 年三月初,均出现南京地区降水量显著小于平均值的现象。

4　结论

基于面向对象的方法论,引入周期表的概念作为数据挖掘基本数据组织方式,分为时空对象状态搜索、时空搜索两个步骤对数据进行处理,其思路:

(1)引入周期表的概念,解决了时空语义与数据操作的一致性问题。

(2)定义了频繁出现的时空状态概念,同时将时空关联规则定义为具有时空拓扑频繁出现的时空状态模式。

(3)利用向下闭合引理,使用 Apriori 算法发现频繁集并进行剪枝,以达到快速抽取状态的目的。

(4)设计了 PRules−Miner,用于从频繁出现的时空状态中挖掘时空关联规则。

时空关联规则挖掘实际上是对海量数据认知的过程。我们将海量数据作为信息的输入,设定一系列的处理规则,得到结果。而这些结果正是与真

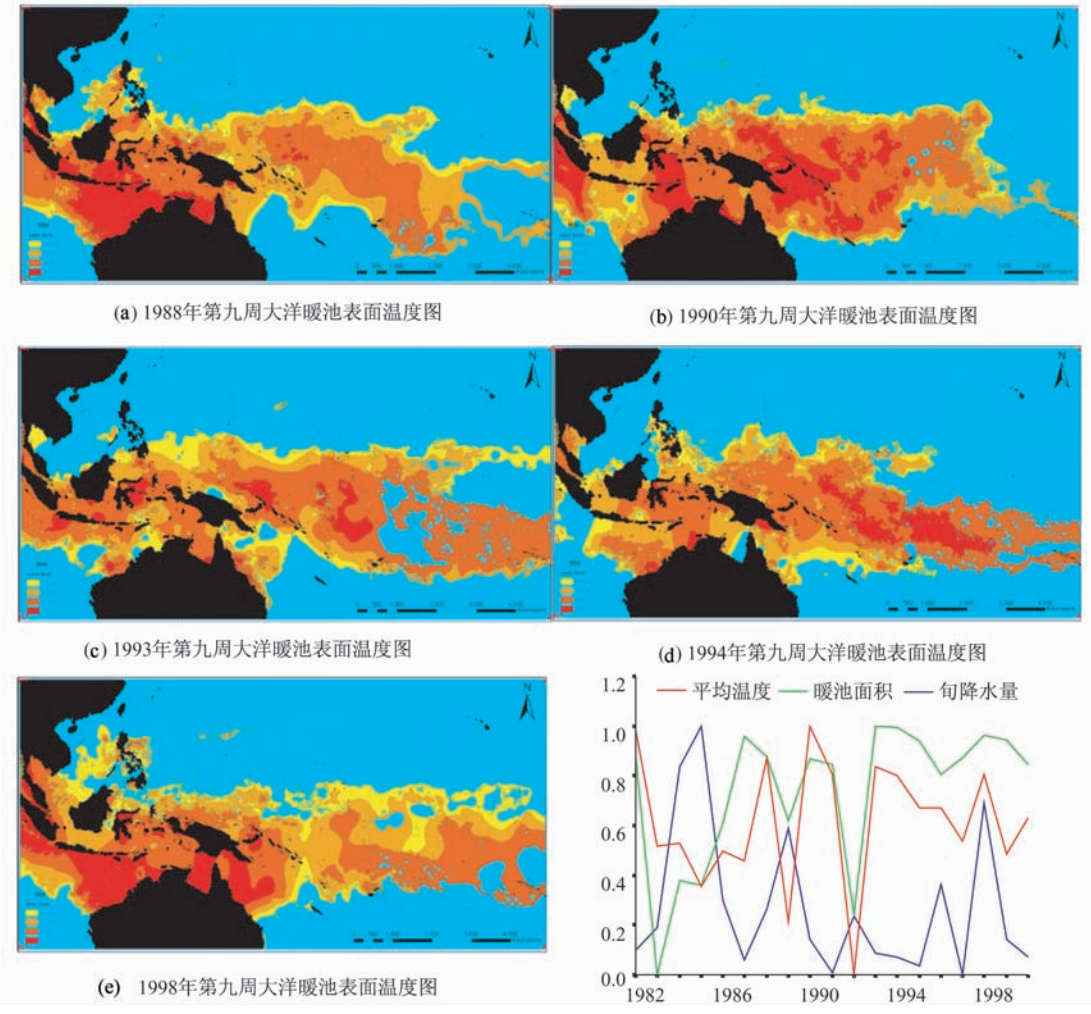

图 4 时空状态状态示例图

Fig. 4 Example charts of Warm Pool status

图 5 大洋暖池属性结合降水属性均一化统计图

Fig. 5 Statistical example chart of spatio-temporal association rule No. 1

实世界中存在的，并可被人类理解的关联规则信息。时空关联信息分段式挖掘时空关联规则的过程是一个逐层归纳、综合信息的过程。实践证明，所得时空关联规则与时空粒度无关，并能够挖掘出时空粒度不一致的地物间的关联关系。算法使用笛卡尔积得到在时空拓扑阈值内匹配的时空候选集，并可以发现时域、空域均不邻接的事物间的时空关联规则，即事物间的“遥相关”。

参考文献：

[1] Verhein F and Chawla S. Mining Spatio-temporal Patterns in Object Mobility Databases[J]. Data Mining and Knowledge Discovery, 2008,16(1): 5 - 38.

[2] Huang Y P, Kao L J and Sandnes F E. Efficient Mining of Salinity and Temperature Association Rules from ARGO Data[J]. Expert Systems with Applications, 2008,35(1 - 2): 59 - 68.

[3] Li Y, *et al*. Discovering Calendar-based Temporal Association Rules[J]. Data & Knowledge Engineering, 2003. 44(2): 193 - 218.

[4] Kalnis P, Mamoulis N and Bakiras S. On Discovering Moving Clusters in Spatio-temporal Data[J]. Advances in Spatial and Temporal Databases, 2005,364 - 381.

[5] Lee A J T, Chen Y A and Weng C C Ip. Mining Frequent Trajectory Patterns in Spatial-temporal Databases [J]. Information Sciences, 2009, 179 (13): 2218 - 2231.

[6] Lee J W, Paek O H and Ryu K H. Temporal Moving Pattern Mining for Location-based Service[J]. Journal of Systems and Software, 2004,73(3): 481 - 490.

[7] Su F, Zhou C, Lyne V, Du Y and Shi W. A Data-mining Approach to Determine the Spatio-temporal Rela-

tionship between Environmental Factors and Fish Distribution[J]. Ecological Modelling, 2004,174(4): 421 - 431.

[8] 孙建奇, 袁薇,高玉中. 阿拉伯半岛-北太平洋型遥相关及其与亚洲夏季风的关系[J]. 中国科学(D 辑:地球科学),2008,38(6): 750 - 762.

[9] 张雪伍, 苏奋振, 石忆邵,等. 空间关联规则挖掘研究进展[J]. 地理科学进展,2007,26(6): 119 - 128.

[10] Agrawal R and Srikant R. Fast Algorithms for Mining Association Rules[A]. 20th Int. Conf. Very Large Data Bases, Santiago de Chile, Chile, Citeseer, 1994.

[11] Han Jiawei and Kamber M. Data Mining: Concepts and Techniques[M]., San Fransisco, CA, USA: Morgan Kaufmann, 2006.

[12] 刘君强,潘云鹤,挖掘空间关联规则的前缀树算法设计与实现[J],中国图象图形学报, A 辑,2003(4): 476 - 480.

[13] Winarko E and Roddick J F. ARMADA——An Algorithm for Discovering Richer Relative Temporal Association Rules from Interval-based Data[J]. Data & Knowledge Engineering, 2007,63(1): 76 - 90.

[14] Chen C H, Hsu W and Lee M L. Discovering Trends and Relationships among Rules[R]. Database and Expert Systems Applications, Proceedings, 2009, 5690: 603 - 610.

[15] Lee Y J, Lee J W, Chai D J, Hwang B H and K. H. Ryu K H. Mining Temporal Interval Relational Rules from Temporal Data[J]. Journal of Systems and Software,2009,82(1): 155 - 167.

[16] Jabas A, Garimella R M, Ramachandram S and Soc I C. MANET Mining: Mining Temporal Association Rules[R]. The 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA - 8) and The 2008 International Conference on Intelligent Pervasive Computing (IPC - 08), 2008.

[17] Beaubouef T, Petry F E and Ladner R. Spatial Data Methods and Vague Regions: A Rough Set Approach [J]. Applied Soft Computing, 2007,7(1): 425 - 440.

[18] Teegavarapu R S V. Estimation of Missing Precipitation Records Integrating Surface Interpolation Techniques and Spatio-temporal Association Rules[J]. Journal of Hydroinformatics, 2009,11(2): 133 - 146.

[19] Zaki M J. SPADE: An Efficient Algorithm for Mining Frequent Sequences[J]. Machine Learning, 2001,42 (1): 31 - 60.

[20] Pujari K A. Data Mining Techniques[M]. Universities Press,2001.

[21] Allen J F. Maintaining Knowledge about Temporal Intervals[J]. Communications of ACM, 1983, 26(11): 832 - 843

[22] Lee W J, Jiang J Y and Lee S J. Mining Fuzzy Periodic Association Rules[J]. Data & Knowledge Engineering,2008,65(3): 442 - 462.

[23] Ozden B, Ramaswamy S and Silberschatz A. Cyclic Association Rules[R] 14th International Conference on Data Engineering, Orlando, FL , USA, 1998.

[24] Bembenik R and Rybiński H. Mining Spatial Association Rules with No Distance Parameter[C] Intelligent Information Processing and Web Mining, 2006,499 - 508.

[25] Appice A, Berardi M, Ceci M and Malerba D. Mining and Filtering Multi-level Spatial Association Rules with ARES[J]. Foundations of Intelligent Systems, 2005, 342 - 353.

[26] Chen J P and Tan X J. Mining Spatial Association Rules with Geostatistics[R]. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 2008.

[27] Qu L L and Chen Y (Eds.). An Algorithm to Improve the Effectiveness of Association Rules Mining[R]. Harbin, Harbin Institute Technology Publishers, 2005.

[28] Kacar E and Cicekli N K. Discovering Fuzzy Spatial Association Rules[R]. Proc. SPIE 4730, 94 (2002). doi:10.1117/12.460216

[29] Mennis J and Liu J. Mining Association Rules in Spatio-temporal Data[R]. Proc. of the 7the Int'l Conf. on Geocomputation, 2003.

[30] Tao Y, Kollios G, Considine F. Li F and Papadias D. Spatio-temporal Aggregation Using Sketches[R]. 20th International Conference on Data Engineering, IEEE, 2004.

[31] Agrawal R, Imieliński T and Swami A. Mining Association Rules between Sets of Items in Large Databases [J]. ACM SIGMOD Record, 1993,22(2): 207 - 216

[32] 薛峰,王会军,何金海. 马斯克林. 高压和澳大利亚高压的年际变化及其对东亚夏季风降水的影响[J]. 科学通报,2003,48(3): 287 - 291.

[33] Wang H and Fan K. Central-north China Precipitation as Reconstructed from the Qing Dynasty: Signal of the Antarctic Atmospheric Oscillation[J]. Geophysical Research Letters, 2005, 32(24): 1 - 4.

[34] Fan K and Wang H. Antarctic Oscillation and the Dust Weather Frequency in North China[J]. Geophysical Research Letters, 2004,31: 1-4.
[35] Tan H, Dillon T, Hadzic F, Chang E and Feng L. (2006). IMB3-Miner: Mining Induced/Embedded Subtrees by Constraining the Level of Embedding[J]. Lecture Notes in Computer Science, 2006, 3918:450-461.
[36] 王秀荣, 王维国, 刘还珠,等. 北京降水特征与西太副高关系的若干统计[J]. 高原气象,2008(4): 822-829.

Period Table Based Spatio-temporal Association Rules Mining

CHAI Siyue[1,2], SU Fenzhen[1], ZHOU Chenghu[1]
(1. *State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101, *China*; 2. *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: As periodical geographical phenomena cover lots of rules, geographic data mining provides a way to find out such rules. In this paper, an algorithm called PRules-Miner is designed based on period table to mine spatio-temporal association rules. Using this mining model, spatio-temporal data were reorganized from sequential dataset to period table set. And spatio-temporal association rules, which describe the teleconnected movement model of two or more objects, can be dug out through three steps: 1) Filtering disorder data in period table: we extract spatio-temporal frequent status in each row and store such status into spatio-temporal frequent item set; 2) Matching objects in the item set based on downward closure lemma and spatio-temporal topology: we match the objects in order to create the spatio-temporal association candidate set; 3) Verifying the candidate set under spatio-temporal topology to find the rules which have to satisfy the spatio-temporal support and spatio-temporal confidence. And the final rules are the spatio-temporal association rules. To check the validation of the algorithm, we use 20 years' AVHRR Product 016, which is sea surface inversion temperature data provided by PO. DAAC and the same period records of Nanjing's daily precipitation provided by National Academy of Meteorological Sciences to mine the teleconnection rules between Eastern Indo Ocean and Western Pacific Ocean Warm Pool and Nanjing's precipitation. The results show, this mining model has the following characteristics: 1) this algorithm is object-orientated and can describe geographical status independently. Thus, the final spatio-temporal association rules are not correlated with spatial scale or temporal scale. 2) The candidate item set is created by Cartesian product, and it can represent complicated spatio-temporal topology between objects. And the spatio-temporal topology can be set manually so as to find the association of none adjacent objects in spatio-temporal dimensions. After setting spatio-temporal topology, spatio-temporal association rules can be mined and validated from candidate set. In the final rules, one object's frequent status is combined with another object's frequent status with given spatio-temporal topology. Thus, the association of objects with uncertain time lag can be extracted.

Key words: data mining; association rules; spatio-temporal data; hierarchical structure; period table