时空关联规则挖掘算法及其在ITS中的应用

夏 英[1,2] 张 俊[2] 王国胤[2]

(西南交通大学信息科学与技术学院 成都 610031)[1] (重庆邮电大学计算机学院 重庆 400065)[2]

摘 要 同时考虑时间和空间约束,能够在分析过程中及时过滤不相关的数据,提高时空关联规则的获取效率。基于这一思路,在频繁项集的产生过程中同时分析数据的时间有效性和空间关联性,提出了Spatio-Temporal Apriori (STApriori)算法。算法首先对时空数据进行时间段划分和空间关联性分析并形成事务表,然后对空间关联的项集进行连接并产生时空关联规则。实验表明了算法的有效性。该算法在智能交通系统(ITS)的应用,可以利用路段间的时空关联规则分析交通拥堵趋势。

关键词 时空约束,关联规则,交通拥堵趋势

中图法分类号 TP393 **文献标识码** A

Spatio-temporal Association Rule Mining Algorithm and its Application in Intelligent Transportation System

XIA Ying[1,2] ZHANG Jun[2] WANG Guo-yin[2]

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)[1]

(School of Computer Science and Technology, Chongqing University of Posts & Telecommunications, Chongqing 400065, China)[2]

Abstract Taking into account the spatial and temporal constraints simultaneously can filter irrelevant data early and improve the efficiency of discovering spatio-temporal association rule. Based on the idea, Spatio-Temporal Apriori (STApriori) algorithm was proposed. It analyzes the time validity and spatial relativity at the same time during the generation of frequency item sets. It classifies the time duration of spatio-temporal data and considers the spatial relationship firstly and generates the transaction table, then performs join operation on spatial-related item sets. Experiments illuminate that the algorithm is well performed. The algorithm is applied in intelligent transportation system to analyze the trend of traffic congestion by identifying spatio-temporal association between road sections.

Keywords Spatio-temporal constraint, Association rule, Trend of traffic congestion

1 引言

随着计算技术、通讯技术、数字存储技术和高速数据获取技术的迅猛发展,在交通、电力、物流、环境监控、工业生产等领域积累了大量与时间和地理空间相关的数据资源。通过研究空间对象随时间的变化规律,发现数据的时空关联规则,分析数据的时空变化趋势并预测未来的时空状态,对于规划建设、指挥调度、应急管理、信息服务等具有重要的应用价值。

时空关联性分析是研究空间对象随时间的变化规律,反映时空数据在时间和空间上的关联性,时空关联规则挖掘作为时空关联性分析的主要方法之一,国内外已有不少学者对其进行了研究或应用。Florian Verhein 和 Sanjay Chawla[1]提出一种在交通高峰区域进行属性约减的时空关联规则算法STAR(Spatio-Temporal Association Rules),其将关联规则扩展到时空领域。李波[2]等基于相关系数矩阵,利用GIS和Qrigin等分析软件,对洪泽湖水质的时空相关性及其时间和空间分布规律进行了研究。岳慧颖[3]提出SKDM(Shi Kong Data Mining)算法,其先按空间位置生成项目-地址对,再综合时间因素发现带有时空约束的关联规则。SheKhar[4]等利用时空关联规则实现交通流量监测。沙宗尧[5]提出时序空间关联规则挖掘方法,并将该方法应用于土地类型变化的时空关联分析中,用以发现土地覆盖演替规律。

以上方法虽然都将时间和空间作为关联规则的约束条件,但是大多以两阶段分别进行分析,而不是同时考虑时空约束。同时考虑时空约束,能够在分析过程中及时过滤不相关的数据,从而提高算法的效率。

文献[3]提出的SKDM算法适用于解决含空间和时间约束的关联规则挖掘问题,该算法首先考虑空间约束,假设时间区间相同,然后将两者的相关有效时间进行推广和归并,得到相应的关联规则。该算法在关联规则经典算法Aprioir算法[6]的基础上,先后考虑了空间和时间的双重约束,较其他时空关联规则挖掘方法更适合于侧重考虑空间和时间因素的时空关联性分析问题。

本文借鉴SKDM算法的思路,在产生频繁项集的过程中

到稿日期:2010-10-11 返修日期:2011-01-05 本文受国家自然科学基金(60773113),重庆市计算机网络与通信技术重点实验室开放基金项目(CY-CNCL-2009-01),重庆市科委科技项目(CSTC2009CB2015)资助。

夏 英(1972-),女,博士生,副教授,主要研究方向是数据库与数据挖掘、地理信息系统等,E-mail:xiaying@cqupt.edu.cn;张 俊(1983-),女,硕士生,主要研究方向为时空数据挖掘、智能交通系统;王国胤(1970-),男,博士,教授,博士生导师,主要研究方向为智能信息处理、信息安全等。

同时考虑时间有效性和空间关联性，提出 Spatio-Temporal Apriori(STApriori) 算法。该算法首先对时空数据进行时间段划分和空间关联性分析并形成事务表，然后对空间关联的项集进行连接并产生时空关联规则。另外，将该算法应用于 ITS，通过分析路段间的邻接关系，结合各路段的交通流量和平均速度，进行交通拥堵趋势分析与预测。

2 时空关联规则的相关概念

2.1 关联规则

设 I 是所有项目的集合，项的集合称为项集，包含 k 个项的项集称为 k 项集[7]。设 T 为事务表，每个事务 T_i 是一个项集，且 $T_i \subset I$。

设 A,B 都是项集，事务 T_i 包含 A 当且仅当 $A \subseteq T_i$。关联规则[7]是形如 $A \Rightarrow B$ 的蕴含式，其中 $A \subset I, B \subset I$，且 $A \cap B = \phi$，A 称为规则的前提，B 称为规则的结果。

规则 $A \Rightarrow B$ 的支持度[7](support)，表明 T 中事务同时包含 A 和 B 的概率，即概率 $P(A \cup B)$。

规则 $A \Rightarrow B$ 的置信度[7](confidence)，表明 T 中包含 A 的事务同时也包含 B 的概率，即条件概率 $P(B/A)$。

项集的出现频率是包含项集的事务数，简称为项集的支持度[7]。如果项集的支持度大于或等于预先定义的最小支持度阈值，则项集是频繁项集[7]，频繁 k 项集的集合通常记作 L_k；候选 k 项集[7]是由频繁 $(k-1)$ 项集与自身连接生成，该候选项集记作 C_k。

同时满足最小支持度(minSup)阈值和最小置信度(minConf)阈值的关联规则称作强关联规则[7]。

给定一个事务集 T，挖掘关联规则问题可以分为两个子问题：(1) 找到所有支持度大于或等于最小支持度(minSup)的项集，这些项集称为频繁项集；(2) 使用第一步找到的频繁项集根据最小置信度产生期望的规则。

Agrawal 等人在 1994 年提出挖掘关联规则的经典算法——Apriori 算法，该算法采用一种被称作逐层搜索的迭代方法，k 项集用于搜索 $(k+1)$ 项集。Apriori-gen() 函数则作为 Apriori 算法中获得频繁项集的关键步骤，备受学者的关注。该函数主要进行两个动作：连接和剪枝，根据频繁 $(k-1)$ 项集自连接生成候选 k 项集，具体过程如文献[6]所述。许多学者在研究 Apriori 算法的基础上，针对 Apriori-gen() 函数的连接步骤进行了改进，如文献[8]提到的 Apriori-like 算法，改进了 Apriori-gen() 函数生成候选项集的过程，扩展了自连接的约束条件，从而获得更多的候选项集；但是该算法扩展的自连接过程是针对项集中单个项本身，没有考虑项集之间的空间关联性，应用于时空关联性问题分析，仍存在一些不足。本文提出的 STApriori 算法中的 Apriori-gen() 函数，其侧重点则在于体现和保证项集之间的空间关联性。

2.2 空间关联性分析

拓扑、距离和方位是 3 种基本的二元空间关联关系，并通过空间谓词[9]描述。如用邻接、关联、包含等谓词表示拓扑关系，用邻近、远离等谓词表示距离关系，用东、西、左、右等谓词表示方位关系。

现实生活中，有些空间对象在距离上是邻近的，但在拓扑上却是不可达的，如被一条江相隔的两条道路。本文结合空间拓扑和距离关系进行空间关联性分析。如在交通路网中，如果上游路口 X 与下游路口 Y 不仅空间上可达且距离邻近，则用空间谓词 close_to(X,Y) 表示，其表示 X 与 Y 是空间关联的。

2.3 时空关联规则

时空关联规则在传统关联规则定义的基础上增加了空间和时间约束。本文定义时空关联规则：

$$P_1 \wedge P_2 \wedge \cdots \wedge P_m \Rightarrow Q_1 \wedge Q_2 \wedge \cdots \wedge Q_n [\text{Valid-Time}], (\text{minSup}, \text{minConf})$$

式中，$P_1, \cdots, P_m, \cdots, Q_1, \cdots, Q_n$ 中至少有一个是空间谓词，Valid-Time 是指有效时间，minSup 是规则的最小支持度，minConf 是规则的最小置信度。

3 STApriori 算法

3.1 算法思路与处理过程

STApriori 算法主要是在 SKDM 算法按空间位置生成项目-地址对的思想基础上，同时考虑时间和空间约束来进行时空关联规则挖掘，具体如图 1 所示。主要的改进表现在以下两个方面：

首先，根据时间段划分和空间关联性分析构造事务表。事务表中的事务按时间段分组，由编号 TID、开始时间戳 starttime、结束时间戳 endtime 和项集 itemset 组成。其中，项集 itemset 包括若干个空间上邻接的(origin,destination)地址对。(origin,destination)地址对能够直观地表达 close_to 谓词。

其次，为保证项集之间的空间关联性，在执行 Apriori-gen() 的 join 操作，由频繁 $(k-1)$ 项集 L_{k-1} 与自身连接产生候选 k 项集时，增加两种约束条件。当 $k=2$，即由频繁 1 项集生成候选 2 项集时，为了保证项集之间的空间关联性，需要比较 L_1 中每个(origin,destination)地址对的两个元素，当某一个地址对的 origin(或 destination)与另一个的 destination(或 origin)相等时才进行连接；当 $k \geq 3$ 时，同样为了保证项集之间的空间关联性，需要对自连接的 L_{k-1} 各项集求交集，当交集的个数为 $k-2$ 时才进行连接。

3.2 算法描述

输入：事务表 T，最小支持度 minSup，最小置信度 minConf。

输出：满足最小支持度和最小置信度的时空关联规则集。

```
STApriori:
第一步产生频繁项集:
(1) L1={frequent 1-itemset} //频繁 1 项集
(2) for(k=2; Lk-1≠φ;k++)
(3)   Ck=Apriori-gen(LK-1);
       //由频繁 k-1 项集生成候选 k 项集
(4)   for each transaction t ∈ T {
       //扫描按时间进行分组的每个事务 t
(5)     Ct=Subset(Ck, t)
         //获取 t 中是候选项集的子集
(6)     for each candidate c∈Ct
(7)           c. Count ++;
          //统计候选 k 项集的计数
(8)     }
(9) Lk= {c∈Ck|c. count >=minSup};
          //计数大于 minSup 的为频繁项集
```

```
(10) return L= ∪k Lk; //返回频繁项集集合
第二步由产生的频繁项集生成关联规则:
(11) L=∪k Lk; // L 是频繁项集集合
(12) AR=ϕ;//AR 是强关联规则集合
(13) for all λk{ //λk 是 L 的元素,是一个频繁 k 项集
(14)   for all ak{// ak 是 λk 的非空子集
(15)     if(ak=>(λk-ak)的置信度>= minConf){
(16)       AR=AR∪ak⇒(λk- ak);
(17)     }
(18)   }
(19) }
(20) return AR;
```

图 1 STApriori 算法

设 I_1,I_2 是长度为 $k-1$ 的频繁项集,V_1,V_2 是两个地址对,Apriori-gen()函数的 join 操作改进如图 2 所示。

```
Apriori-gen():
join:
  if (k=2)
    if (I1. V1. origin=I2. V2. destination and I1. V1. destination
      ≠I2. V2. origin) or (I1. V1. destination= I2. V2. origin
      and I1. V1. origin≠I2. V2. destination)
    Ck=Ck∪{I1, I2}
  else
    if((I1∩I2). size==k-2)
   Ck=Ck∪{I1∪(I2- I1)}
  prune:
    For all itemset c ∈Ck do
      for all (k-1)-subsets s of c do
        if (s is not in Lk-1) then
      Delete c from Ck
```

图 2 Apriori-gen() 函数

3.3 算法分析与结果

相对于 SKDM 算法,本文提出的 STApriori 算法同时考虑了空间对象之间的空间拓扑和距离关系,并同时进行空间关联性和时间有效性分析。将这两个算法在同一测试数据集上进行分析对比。

实验使用的环境为 Genuine Intel(R) CPU 2140 @ 1.60GHz 1.00GB RAM,Microsoft Windows XP Professional,算法使用 Eclipse 平台的 Java 语言实现,数据集由 dynaCHINA[10] 产生。实验分两种情况进行:

1)测试数据集相同,在不同的最小支持度条件下对比算法的执行时间。实验结果表明,STApriori 算法的执行时间明显优于 SKDM 算法,且性能稳定,如图 3 所示。这是由于 STApriori 算法同时考虑了空间对象之间的空间拓扑和距离关系,并同时进行空间关联性和时间有效性分析,及时过滤了不相关的数据,减小了计算量。

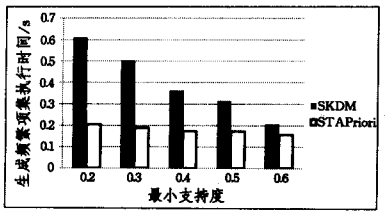

图 3 不同最小支持度下的算法执行时间

2)保持最小支持度不变(如 0.3),对 5 种不同的测试数据集进行实验,使用的测试数据集分别由 dynaCHINA[10] 产生,比较生成频繁项集执行时间的结果如图 4 所示。

实验结果表明,随着数据集中数据量的增加,由于 STApriori 算法同时考虑了时间和空间因素,及时过滤了不相关的数据,因此在时间效率上明显优于 SKDM 算法;而 SKDM 算法由于先考虑空间因素,后考虑时间因素,对于路网分布比较复杂的交通流量数据集(如图中的第 3,4,5 种数据集),在生成项目-地址对的过程中浪费大量的时间,因此时间效率明显低于 STApriori 算法。

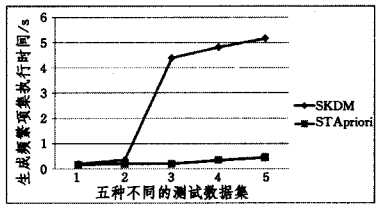

图 4 不同测试数据集下的算法执行时间(最小支持度 0.3)

算法效率的两个度量指标是:算法的执行时间和算法的存储空间需求。由图 3 和图 4 的结果可以得知本算法的执行时间明显优于 SKDM 算法;而从算法的存储空间需求上来讲,本算法主要改进了 Apriori-gen()的 join 操作,在由频繁($k-1$)项集 L_{k-1} 与自身连接产生候选 k 项集时,增加了空间关联性的约束条件,从而及时过滤了大量空间上无关联的数据,减少了产生候选项集对存储空间的需求,最终降低了算法对存储空间的需求。

4 实例分析

将该时空关联性分析应用于 ITS 中,根据交通流量信息首先进行路段间的空间关联性分析,再利用 STApriori 算法生成时空关联规则,可对某区域和时间段内的交通拥堵进行趋势分析与预测,有利于道路导航、交通控制等应用。具体处理过程如图 5 所示。

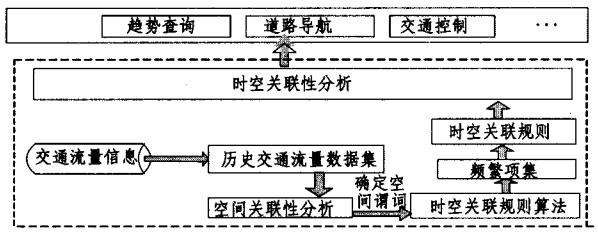

图 5 时空关联分析在 ITS 中的应用

采用 dynaCHINA 数据发生器,基于路网产生 7:15-8:45 的交通流量数据集,每个数据项包括上游路口、下游路口、交通流量、车辆平均速度、开始时间、终止时间等 6 个字段,共 156 个数据项。

本文将交通流量与车辆平均速度相结合进行交通拥堵状态的判断,即根据领域专家意见设定交通流量阈值和车辆平均速度阈值,大于设定的流量阈值并且小于设定的平均速度阈值则认为对应的交通状态为拥堵。对判断为拥堵的数据项,按相同开始时间戳和结束时间戳分组,并用 close_to 谓词描述空间上邻接的上游和下游路口,从而构成一个事务并插入到如表 1 所列的事务表中。与普通的事务数据表不同,本文的事务数据表不仅包含事务号 ID、事务项(origin,destination)对,还包括 starttime 和 endtime 项。以 15 分钟为周期进

行交通拥堵状况的分析和判断，整个事务表总共包括如下 6 个时间分组：7:15:00－7:30:00，7:30:00－7:45:00，7:45:00－8:00:00，8:00:00－8:15:00，8:15:00-8:30:00，8:30:00－8:45:00，对其进行算法分析。

表 1　事务表

TID	itemset	starttime	endtime
T_1	close_to(9,1) close_to(19,4) close_to(6,7)	7:15:00	7:30:00
T_2	close_to(9,1) close_to(1,2) close_to(11,2) close_to(2,3) close_to(3,4) close_to(19,4) close_to(19,20)	7:30:00	7:45:00
T_3	close_to(5,6) close_to(6,7) close_to(9,1) close_to(1,2) close_to(11,2) …	7:45:00	8:00:00
…	…	…	…

本文以 minSup 为 0.50，minConf 为 0.65 为例进行分析，首先得到频繁 1 项集 L_1，如表 2 所列。由 L_1 自身连接产生候选 2 项集 C_2，并记录生成 C_2 所扫描的每个 T_i，所得的候选 2 项集 C_2 如表 3 所列，将表 3 所得的 C_2 与支持阈值进行比较，大于 minSup 的作为频繁 2 项集 L_2，生成表 4 的结果。当 L_2 自身连接结果为空，即 $C_3=\phi$ 时，算法终止，至此找出了所有的频繁项集。

表 2　频繁 1 项集

itemset	support
[9,1]	1.0
[19,4]	1.0
[1,2]	0.67
[11,2]	0.83
[3,4]	0.5
[12,3]	0.5
[4,3]	0.67
[15,5]	0.5

表 3　候选 2 项集的集合

itemset	support	TID
[1,2],[9,1]	0.67	T_2,T_3,T_4,T_5
[4,3],[19,4]	0.67	T_3,T_4,T_5,T_6
[12,3],[3,4]	0.33	T_3,T_4

表 4　频繁 2 项集

itemset	support	TID
[1,2],[9,1]	0.67	T_2,T_3,T_4,T_5
[4,3],[19,4]	0.67	T_3,T_4,T_5,T_6

根据设定的 minConf 阈值，由频繁项集生成的关联规则如表 5 所列，所得的关联规则均属于强关联规则。

表 5　关联规则表

规则	支持度	置信度	有效时间
[1,2]⇒[9,1]	0.67	1.0	7:30:00－8:30:00
[9,1]⇒[1,2]	0.67	0.67	7:30:00－8:30:00
[4,3]⇒[19,4]	0.67	1.0	7:45:00－8:45:00
[19,4]⇒[4,3]	0.67	0.67	7:45:00－8:45:00

根据上述得到的关联规则表，以[1,2]⇒[9,1]规则为例分析如下：

规则[1,2]⇒[9,1]，支持度为 0.67，置信度为 1.0，表明在时间段 7:30:00 到 8:30:00 内，如果以 1 为 origin、以 2 为 destination 的路段发生拥堵，那么以 9 为 origin、以 1 为 destination 的路段发生拥堵的可能性是 0.67，在事务表中该规则的置信度是 1.0。结合 close_to 谓词，可以得到时空关联规则如下：

is_a(1,origin) ∧ is_a(2,destination) ∧ close_to(1,2) ∧ congestion(1,2)⇒is_a(9,origin) ∧ is_a(1,destination) ∧ close_to(9,1) ∧ congestion(9,1),[7:30:00,8:30:00],(0.67,1.0)

交通拥堵按照形成的先后次序可分为初始交通拥堵和后续交通拥堵。结合路网，根据此规则可以判断路段[1,2]是初始交通拥堵路段，且可以推断出在时间段 7:30:00 到 8:30:00 内该路段拥堵会导致路段[9,1]的后续交通拥堵，这种情况出现的概率是 0.67。

规则[9,1]⇒[1,2]与规则[1,2]⇒[9,1]的不同之处在于，结合路网可以判断[9,1]是后续交通拥堵路段，由此可以推断出在时间段 7:30:00 到 8:30:00 内导致该路段拥堵的起始交通拥堵路段是[1,2]，这种情况出现的概率是 0.67。

结束语　本文在 SKDM 算法基础上，提出了一种时空关联规则算法 STApriori。该算法同时考虑了时间的有效性和空间的关联性。通过实验对比分析证明该算法的正确性和有效性，将该算法应用于 ITS 中，进行交通拥堵的趋势分析与预测，可以分析造成后续拥堵的原因，也可以预测初始拥堵会造成的影响。目前，算法研究主要是在静态历史数据的基础上开展的，下一步将考虑对实时交通数据的处理和分析。

参 考 文 献

[1] Verhein F, Chawla S. Mining Spatio-temporal patterns in object mobility database[C]// Data Mining and Knowledge Discovery. Hingham, Kluwer Academic Publishers, 2008: 5-38

[2] 李波，濮培民，韩爱民. 洪泽湖水质的时空相关性分析[J]. 湖泊科学，2001，14(3)：259-266

[3] 岳慧颖. 含有时空约束的关联规则挖掘方法研究[D]. 哈尔滨：哈尔滨工程大学，2004(4)

[4] Shekhar S, Huang Y. Discovering spatial co-location patterns: a summary of results[C]// Lecture Notes in Computer Science. Berlin Heidelberg, Springer-Verlag, 2001: 236-256

[5] 沙宗尧. 时序空间关联规则挖掘及其应用研究[J]. 地理空间信息，2008，6(5)：18-21

[6] Agrawl R, Srikant R. Fast algorithms for mining association rules[C]// Proceedings of the 20th VLDB Conference. Santiago Chile, 1994

[7] Han Jia-wei, Kamber M. 数据挖掘概念与技术(第二版)[M]. 范明，孟小峰，译. 北京：机械工业出版社，2006：148-151

[8] Gorawski M, Jureczek P. Using Apriori-like Algorithms for Spatio-Temporal Pattern Queries[C]// Proceedings of the International Multiconference on Computer Science and Information Technology. Poland, Polish Information Processing Society, 2009: 43-48

[9] 许红，严静，张群洪. 基于概念树的空间关联规则挖掘算法及其在土地利用分析中的应用[J]. 华中农业大学学报：社会科学版，2004(6)：46-50

[10] 许兆霞，林勇，李树斌，等. 实时交通预估预测仿真系统 dynaCHINA参数标定及应用[J]. 山东科学，2009，22(6)：46-49