

# 基于 FP-树的时空关联规则挖掘算法研究

张雪萍,李围成,祝玉华

(河南工业大学 信息科学与工程学院,河南 郑州 450001)

**摘 要:** 提出一种基于 FP-树的时空关联规则挖掘算法,即 STFP-tree(Spatio-temporal FP-tree)算法. 算法对时空数据进行时间上的划分,并用空间谓词分析空间对象间的位置关系,最终生成同时考虑时间和空间特征的 FP-树,进而挖掘出强时空关联规则. 实验结果证明了算法的有效性,适用于对时空数据进行分析挖掘.

**关键词:** 时空数据;FP-树;挖掘算法;位置关系;时空关联规则

**中图分类号:** TP301

**文献标识码:** A

**文章编号:** 1000-7180(2016)08-0130-04

## Research on Spatio-temporal Association Rules Mining Algorithm Based on FP-tree

ZHANG Xue-ping, LI Wei-cheng, ZHU Yu-hua

(School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China)

**Abstract:** A spatio-temporal association rules mining algorithm based on FP-tree(STFP-tree) is proposed. It divides the data into time, and uses spatial predicates to analyze the position relationship between the spatial objects. Finally, it generates a FP-tree, which considers spatio-temporal features, and then gets strong spatio-temporal association rules. Experiments demonstrate that the algorithm can be well performed. It is suitable for the analysis and mining of spatio-temporal data.

**Key words:** spatio-temporal data; FP-tree; mining algorithm; position relationship; spatio-temporal association rules

### 1 引言

时空关联规则挖掘是对关联规则挖掘的一个扩展,是时空数据挖掘的一种主要方法. 时空关联规则挖掘主要是在传统关联规则挖掘中添加时间和空间约束,对时空数据进行分析,找出包含时间、空间以及非空间属性的强关联规则,从而发现存在的潜在规律和联系.

国内外学者对于时空数据挖掘已经做出了很多有价值的研究,文献[1]提出使用基于周期表的时空关联规则挖掘方法<sup>[1]</sup>发现存在一定距离但仍然存在相关关系的一类空间对象. 文献[2]定义了一个 STARs(Spatio-Temporal Association Rules)规则,用于对某段时间内区域间对象移动的方式进行预测分析并提出了 STAR-Miner 算法,用于寻找符合

STARs 规则的强关联模式. 定义了一种 SOAP (spatial object association patterns)规则<sup>[3]</sup>,并将其用于挖掘生物信息科学、计算流体动力学、分子动力学等领域的时空关联模式. Hamed Chok 和 Le Gruenwald 研究出了一种在线时空关联规则挖掘的框架,并将其应用于传感器网络中<sup>[4]</sup>. Du Xiaoxi 等提出了一种时空关联规则算法并将其用于金融市场中,以发现金融市场中的迁移主题<sup>[5]</sup>. Leong Kelvin 和 Li Junco 等通过提出一种三层的动态模型分析框架,并将其应用在了用于对犯罪的预防中<sup>[6]</sup>. 夏英、张俊和王国胤提出了一种改进的 Apriori 算法,并应用于智能交通系统中,实现交通拥堵趋势分析<sup>[7]</sup>.

本文提出的 STFP-tree 算法,其设计思想主要借鉴了经典的 FP-growth 算法<sup>[8]</sup>,对时空数据进行

时间上的划分,并用空间谓词分析空间数据间的位置关系,将时间和空间特征转移到生成 FP-树的过程中并确定频繁项集,最终提取出强时空关联规则,然后结合实例对所在学校周围酒店住宿价格进行趋势的分析和预测。

## 2 时空关联规则挖掘的基本概念

### 2.1 关联规则挖掘的概念

关联规则挖掘的目标是通过一定的方法,在数据库中寻找出数据项之间的隐藏的某种关联关系,并形成相应的规则,其理论最早是由 Agrawal 等人提出,他们还提出了经典的 Apriori 算法<sup>[9]</sup>。

设  $I = \{i_1, i_2, i_3, i_4, \dots, i_n\}$  是数据库中所有项的集合,  $T = \{t_1, t_2, t_3, t_4, \dots, t_m\}$  是数据库中所有事务的集合,数据库中的项组成数据库中的各个事务。设  $0 \leq k \leq n$ ,则拥有  $k$  个项的集合称为  $k$ -项集。设  $X, Y$  为  $I$  中两类互不相交的数据项集,用公式表示就是:  $\{X \subset I, Y \subset I \mid X \cap Y = \emptyset\}$ ,则关联规则是以  $X \rightarrow Y$  这样的形式表示数据项之间存在某种联系。

关联规则可采用支持度 (Support) 和置信度 (Confidence) 来对其关联的强度进行度量。支持度用于表示  $X$  和  $Y$  同时出现的频繁程度,置信度则用于表示  $Y$  中包含  $X$  的事务出现的频繁程度。支持度 (S) 和置信度 (C) 度量可以用如下公式表达<sup>[10]</sup>:

$$S(X \rightarrow Y) = p(X \cup Y) \quad (1)$$

$$C(X \rightarrow Y) = P(Y/X) \quad (2)$$

关联规则的提取,是要找出支持度和置信度都大于等于预先定义好的最小支持度 (minSupport) 和最小置信度 (minConfidence) 的所有强关联规则。

### 2.2 空间关联规则挖掘概念

空间关联规则挖掘基于关联规则挖掘的思想,主要是把关联规则的理论方法用于对含有空间特征的空间数据进行挖掘,从而发现数据对象之间的空间联系规则。Koperski Krzysztof 和 Han Jiawei 最早将关联规则应用在空间数据库中<sup>[11]</sup>。空间关联规则在传统关联规则理论的基础上做出了部分改进,主要是把传统的数据项变成了空间谓词,其变化的表达式如下:

$$X_1 \wedge X_2 \wedge \dots \wedge X_n \rightarrow Y_1 \wedge Y_2 \wedge \dots \wedge Y_m \quad (3)$$

式中设  $X = (X_1, X_2 \dots X_n), Y = (Y_1, Y_2 \dots Y_m)$ ,  $X$  和  $Y$  分别是  $X_i$  和  $Y_j$  的空间谓词集合,  $X_i$  和  $Y_j$  表示的是空间谓词或非空间谓词,而且其中至少要包含一

个空间谓词并且  $X_i$  与  $Y_j$  的交集为空。关联程度仍然使用支持度和置信度来进行度量,支持度 (S),表示  $X$  和  $Y$  在空间事务中同时发生的频繁程度,记为  $P(X \cup Y)$ 。置信度 (C),表示在空间事务中  $B$  中包含  $A$  的空间事务出现的频繁程度,记为  $p(X/Y)$ 。

### 2.3 时空关联规则挖掘概念

时空关联规则是在关联规则和空间关联规则的基础上提出的,其理论方法主要是针对于含有时间和空间约束的数据进行分析挖掘,在空间关联规则的基础上添加了时间上的约束。其表示形式可以定义<sup>[11]</sup>为

$$X_1 \wedge X_2 \wedge \dots \wedge X_n \rightarrow Y_1 \wedge Y_2 \wedge \dots \wedge Y_m \quad (\text{Time-constrain}) \quad (4)$$

式中,  $X = (X_1, X_2 \dots X_n), Y = (Y_1, Y_2 \dots Y_m)$ ,  $\text{Time-constrain} = (T_1, T_2 \dots T_n)$ 。  $T_k$  表示各个时间点,  $X$  和  $Y$  分别是  $X_i$  和  $Y_j$  的空间谓词集合。  $X_i$  和  $Y_j$  除了要满足其中至少要包含一个空间谓词并且  $X_i$  与  $Y_j$  的交集为空外,还需要标识其具体的产生时间  $T_k$ 。支持度 (S) 则表示的是在某一时间约束下,  $X$  和  $Y$  在空间事务中同时发生的频繁程度,记为  $P(X \cup Y)(\text{Time-constrain})$ 。置信度 (C) 也是在其时间约束下,表示在空间事务中  $B$  中包含  $A$  的空间事务出现的频繁程度,记为:  $p(X/Y)(\text{Time-constrain})$ 。

时空关联规则挖掘是要提取出所有符合一定时间约束条件下的强空间关联规则,时空关联规则不仅是时间约束下的频繁项集也是空间约束下的频繁项集,是时间频繁项集与空间频繁项集之间的交集。

## 3 STFP-tree 算法及实现过程

STFP-tree 算法同时考虑了时间和空间特征并对强时空关联规则进行提取,其功能是直接通过 FP-树找出频繁项集。主要改变是在对空间数据集进行关联规则提取时,需要先将空间对象转化为用空间谓词表达的空间关系,然后把空间关系作为数据项进行处理,同时需要记录目标时间内各个事务发生的时间点,将时间点也泛化成数据项进行分析。这样在对频繁项集进行提取时,就提取出了同时满足支持度和置信度阈值的时间点、空间位置关系以及普通属性三种数据项的频繁项集,从而发现其强时空关联规则。基于 STFP-tree 算法的时空关联规则挖掘的具体过程包含以下阶段。

第一个阶段是对收集到的时空数据进行数据的转换处理,以使得数据可以更好的进行处理,主要包括以下几个步骤。

(1)首先对采集的时空数据进行数据预处理. 过滤掉不相关的冗余数据, 包括不在研究目标时间范围内的数据, 不在研究目前区域内的数据和存在空间依赖的数据.

(2)然后对空间数据进行形式化处理, 将空间对象按空间谓词对空间关系进行合理划分, 具体是以研究对象为中心, 以周围其他对象为目标, 形成<空间谓词, 目标对象>的空间关联关系.

(3)接着按事务具体发生的时间点, 生成事务表来存储处理过的时空数据, 事务表主要包括以下几个字段: 事务编号 ID、空间位置关系、对象属性、发生时间点.

第二阶段是用 STFP-tree 算法对第一阶段生成的事务表进行处理分析, 生成强时空关联规则, 最后再结合相关背景知识对规则进行解释和说明. 具体 STFP-tree 算法的描述如下所示.

(1)初始化: 把第一阶段生成的事务表作为输入数据, 确定最小支持度和最小置信度.

(2)第一次扫描事务表, 去除掉事务表中所有支持度数小于最小支持度的数据项, 记录剩余数据项的支持度数并按降序进行排列, 生成频繁 1-项集.

(3)第二次扫描事务表, 并对每条事务信息按照频繁 1-项集的顺序进行排序.

(4)初始化 FP-树, 以字符等于“NULL”为根节点, 按照事务表顺序依次插入每条事务数据, 生成对应的分支路径同时建立项头表. 对于加入的数据节点需要在其后面统计其出现的次数, 形式为  $(X_i, N)$ ,  $X_i$  为输入的项,  $N$  为出现的次数,  $N$  初始化为 1. 对于要插入的每条事务记录, 其添加方式分为两种情况, 第一种情况是如果与之前已经插入的事务记录有共同的前缀, 则合并共同的前缀项并将每个前缀项的出现次数  $N$  更新为  $N+1$ , 然后为余下不同的数据项添加新的分支并记录  $N$  为 1. 第二种情况是如果与之前已经插入的事务记录没有共同前缀, 则在根节点建立新的分支同时记录  $N$  为 1. 项头表分别存储了数据项、支持度数和头节点指针, 头节点指针指向由相同数据项组成的节点链的第一项的位置.

(5)从 FP-树中找出频繁项: 从最小支持度数的数据项开始按顺序遍历项头表中的每一项, 依次生成各个节点的条件模式基和条件模式树. 条件模式基是由各条包含各个节点的前缀子路径组成. 条件模式树是基于条件模式基构建的, 构建条件模式树的过程是一个递归的树的修剪过程, 需删除不满足

最小支持度计数的节点, 最终生成包含各个数据项节点的频繁项集.

4 实验结果与分析比较

对于提出的 STFP-tree 算法, 结合实例来说明算法执行的有效性. 酒店产业作为社会经济中一条庞大的经济链, 其影响越来越广. 酒店的住宿价格、所在的地理位置和所处的时间阶段都影响着酒店的整体发展前景. 所以, 通过对实地酒店位置的考察, 模拟了郑州市的 3 所学校及其附近 1 000 m 内 7 所酒店的空间分布信息.

采用提出的基于 STFP-tree 算法的操作步骤对时空数据进行时空关联规则的提取, 首先计算酒店到学校的距离信息, 取最短的距离作为最终确定酒店与学校空间位置关系的度量条件, 形成空间关系表. 按照酒店住宿平均价格将酒店的价格进行等级划分, 形成酒店价格等级表. 把酒店价格等级和酒店与学校的位置关系进行归并, 最终转换成了包括时间和空间信息的时空数据事务表.

将时空事务表中的空间关系、价格等级、交易时间作为数据项, 进行时空关联规则的提取. 定义最小支持度数(minSupport)为 3, 首先去除掉事务表中所有支持度数小于 minSupport 的数据项, 生成频繁 1-项集并按支持度数降序排列.

以字符等于“NULL”为根节点, 构建 FP-树, 同时按照事务表和频繁 1-项集的顺序依次插入每条事务数据, 生成对应的分支路径同时建立项头表, FP-树和项头表如图 1 所示.

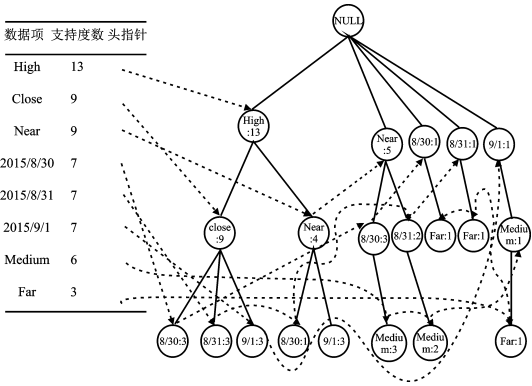


图 1 FP-树与项头表

接着从 FP-树中结合项头表采用递归的方式从最小支持度数的数据项开始生产各个数据项的条件模式基和条件模式树, 从而提取出频繁项集.

对提取的频繁项集进行分析, 计算其置信度, 即

可得出强时空关联规则,规定最小置信度(minConfence)为 0.4. 下面分俩种情况来对关联规则进行提取.

(1)空间关联规则分析:由频繁项集生成的空间关联规则如表 1 所示.

| 表 1 空间关联规则  |      |      |
|-------------|------|------|
| 空间规则        | 支持度数 | 置信度  |
| Near→Medium | 5    | 0.56 |
| Close→High  | 9    | 1    |
| Near→High   | 4    | 0.44 |

结合实验的现实背景知识,可以从空间关联规则表中提取出具有实际意义的空间关联规则表,如表 2 所示.

| 表 2 实际意义的空间关联规则表 |                                                                                    |  |
|------------------|------------------------------------------------------------------------------------|--|
| 编号 ID            | 空间关联规则                                                                             |  |
| 1                | $Is\_a(X,a\_hotel) \wedge Near\_by(X,a\_shool) \rightarrow is\_Medium(X) (5,0.56)$ |  |
| 2                | $Is\_a(X,a\_hotel) \wedge Close\_to(X,a\_shool) \rightarrow is\_High(X) (9,1)$     |  |
| 3                | $Is\_a(X,a\_hotel) \wedge Near\_by(X,a\_shool) \rightarrow is\_High(X) (4,0.44)$   |  |

从表 2 可以发现,酒店与学校的位置关系影响着酒店的住宿价格.酒店与学校的距离影响着酒店的价格定位,并且价格随着距离的增长呈反比.

(2)在空间关联规则中添加时间点约束,得出时空关联规则.

通过对空间关联规则和时空关联规则的分析比较,可以看出它们都可以找出学校附近的酒店的价格随着距离的靠近,其价格也呈现上升趋势的规律.但是通过对时空关联规则的分析,还可以研究出随着时间的不断推移,酒店价格随着时间的变化的波动规律.所以对于酒店价格的分析,需要同时考虑时间和空间的双重影响,这样才能得出可靠的结论,从而对酒店的发展做出合理规划.

## 5 结束语

本文提出了一种针对时空数据挖掘的 STFP-tree 算法.算法在实现过程中把含有空间特征和时间特征的数据存储在特殊的 FP-树中,从而实现了 对数据时空关联规则的提取.将算法应用于对学校周围酒店住宿价格趋势的分析中,实验结果表明了该算法对时空关联规则提取的正确性与有效性,不仅可以发现空间关联规律还能够挖掘出事物随时间

的变化规律,为酒店以后的发展决策提供可靠的指导依据.

## 参考文献:

[1] 柴思跃,苏奋振,周成虎.基于周期表的时空关联规则挖掘方法与实验[J].地球信息科学学报,2011,13(4):455-464.

[2] Verhein Florian, Chawla Sanjay. Mining spatio-temporal patterns in object mobility databases[J]. Data Mining and Knowledge Discovery, 2008, 16(1): 5-38.

[3] Hui Yang, Srinivasan Parthasarathy. Mining spatial and spatio-temporal patterns in scientific data[C]//The 22nd International Conference on Data Engineering. Atlanta: IEEE computer society, 2006.

[4] Hamed C, Le G. An online spatio-temporal association rule mining framework for analyzing and estimating sensor data[C]//Proc of the 2009 Int Database Engineering & Applications Symp. New York:ACM,2009:217-226.

[5] Xiaoxi Du , Ruoming Jin , Liang Ding Victor , et al. Migration Motif: A Spatial-Temporal Pattern Mining Approach for Financial Markets[C]//Proc of the 15th ACM SIGKDD Int Conf on Knowledge and Data Mining. New York:ACM, 2009: 1135-1144.

[6] Kelvin Leong, Junco Li, Stephen Chan, et al. An Application of the Dynamic Pattern Analysis Framework to the Analysis of Spatial-Temporal Crime Relationships[J]. Journal of Universal Computer Scintence, 2009,15(9):1852-1870.

[7] 夏英,张俊,王国胤.时空关联规则挖掘算法及其在ITS中的应用[J].计算机科学,2011,38(9):173-176.

[8] Jiawei Han, Jian Pei, Yiwen Yin. Mining frequent patterns without candidate generation[C]//Proc of the 2000 ACM-SIGMOD Intl Conf on Management of Data. Dallas,TX:ACM Press, 2000:1-12.

[9] Rakesh Agrawal,Ramakrish Srikant. Fast algorithms for mining association rules[C]//Proc of 1994 Int'l Conf on Very Large Data Bases. San-tiago, Chili: VLDB Endowment,1994:487-499.

[10] TAN Pang-ning, STEINBACH Michael, KUMAR Vipin. 数据挖掘导论[M]. 范明,范宏建,译. 北京:人民邮电出版社,2006.

[11] Krzysztof Koperski, Jiawei Han. Discovery of spatial association rules in geographic information databases [C]//prceedings of the 4th International Symposium. Portland,UE,USA:IEEE,1995:47-66

由表 1 数据可知,本文方法的平均相对误差、均方根误差与 K-均值 RBF 方法及 SVM 方法相比较明显降低,说明本文方法的预警准确性很高.

表 1 预警准确性比较

|        | 本文方法 | K-均值<br>RBF 方法 | SVM 方法 |
|--------|------|----------------|--------|
| MAPE/个 | 0.98 | 1.62           | 2.31   |
| RSME/个 | 0.48 | 2.59           | 4.85   |

4 结束语

本文提出一种新的高可用性网络入侵预警方法,介绍了隐马尔可夫模型,在此基础上,构建依据 HMM 模型的实时高可用性网络入侵预警模型,依据危险理论对评估结果进行分析,将高可用性网络状态划分成四个等级,通过构建的 HMM-NSSP 模型实现高可用性网络入侵预警. 为了提高 HMM-NSSP 模型的预警精度,引入最大熵法对高可用性网络安全态势进行一致性判定. 实验结果表明,所提方法预警结果和实际结果基本一致,具有很高的预警精度,可用性很高.

参考文献:

[1] 宋杰,刘雪冰,朱志良,等. 一种能效优化的 MapReduce 资源比模型[J]. 计算机学报,2014,38(1):61-78.

[2] 吕海燕. 网络安全监控数据流的查询优化方法[J]. 电子设计工程,2014,22(13):29-31.

[3] 李绍晖,刘纪伟. 一种面向工业控制系统的改进 CUSUM 入侵检测方法[J]. 电子技术应用,2014,41(9):109-115.

[4] 黄伟建,周鸣爱. MapReduce 高可用性的研究与优化[J]. 计算机工程与设计,2015,35(12):3069-3098.

[5] 左毅,陈勇,游华. 网络入侵应急模型快速响应方法研究与仿真[J]. 计算机仿真,2015,31(3):268-302.

[6] 姜海洋,谢高岗. 众核处理器上的高可用性网络入侵检测系统[J]. 高技术通讯,2015,24(3):859-871.

[7] 黄亮,姜帆,荀浩,等. 面向软件定义网络的入侵容忍控制器架构及实现[J]. 计算机应用,2014,35(11):2898-3024.

[8] 赵夫群. 基于混合核函数的 LSSVM 网络入侵检测方法[J]. 现代电子技术,2014,38(21):87-88.

[9] 杨飞,朱志祥,梁小江. 基于 Ceph 对象存储集群的高可用设计与实现[J]. 微电子学与计算机,2015,32(12):59-63.

[10] 汪波,聂晓伟. 基于多目标数学规划的网络入侵检测方法[J]. 计算机研究与发展,2014,52(10):2340-2367.

作者简介:

龚健虎 男,(1967-),博士,讲师. 研究方向为数据库与知识库和数据挖掘. E-mail:gongtiger@21cn.com

(上接第 133 页)

作者简介:

张雪萍 女,(1968-),博士,教授. 研究方向为空间数据挖掘,智能信息处理技术等.

李围成(通讯作者) 男,(1991-),硕士研究生. 研究方向为空间数据挖掘. E-mail:13598591208@163.com.

祝玉华 女,(1965-),博士,教授. 研究方向为空间数据挖掘,粮食信息处理等.