文章编号:1000-5862(2013)04-0371-05

序列模式挖掘在警用车辆维修数据分析中的研究与应用

滕少华,洪嘉铭,张　巍
(广东工业大学计算机学院,广东 广州　510006)

摘要:将序列模式挖掘方法应用于警用车辆维修数据分析中,对车辆维修记录序列中的多个维度属性展开分析,获取辅助决策信息. 实验结果表明:该序列模式挖掘方法在警用车辆维修数据分析中是可行、有效的.

关键词:序列模式挖掘;警用车辆维修;数据挖掘
中图分类号:TP 391　　**文献标志码**:A

0　引言

近年来,随着信息技术的发展,电子信息技术已深入各企事业单位的方方面面. 信息技术的应用导致企事业单位的运营管理数据呈几何级数增长,面对日益庞大的数据,如何挖掘隐藏在数据后的信息,辅助管理人员进行有效管理决策已成为亟待解决的问题. 迄今,人们已开展了许多数据挖掘技术研究与应用的探索. 许多研究成果已应用于电信、金融、电子商务、电子政务等领域中,然而对警用车辆维修数据的分析,尚处在起步阶段.

车辆维修有其特殊性,一辆车进行了某项维修后,由于服务的保质要求,短期内无需进行同样的维修服务. 但对于警用车辆维修,长期的维修管理数据揭示出较严重的问题,比如:部分车辆维修费用高、维修次数多,短期内部分车辆存在多次维修同种问题的情况等.

针对这些问题,某市交警支队开发了一个警用车辆维修管理系统,开展了对警用车辆维修数据的分析. 通过近 3 年的实践,该系统的应用已为政府节省了大量经费开销,取得了较好的社会效益与经济效益.

序列是事件的有序列表[1]. 对于时间序列数据,其序列数据由相等时间间隔记录的数值数据的长序列组成. 时间序列数据包含不同时间点重复测量得到的数值序列,它可以被许多自然或经济过程产生. 序列模式挖掘关注时间序列模式,序列模式是一个存在于单个序列或一个序列集中的频繁子序列. 序列模式挖掘就是挖掘相对时间或其他模式中出现频率相对高的模式,它已有广泛应用与研究. 例如,将序列模式挖掘应用于网络入侵检测中[2-3],将序列模式应用于音乐类型分类中[4]. 本文细致地分析了警用车辆维修数据,经研究,符合序列模式挖掘要求,因而本文将序列模式挖掘应用于警用车辆维修管理系统,期望发现车辆维修事件中存在的序列模式或规则,用于建立警用车辆维修事件规则库.

1　序列模式挖掘

1.1　序列模式挖掘相关定义

定义 1[5]　项(item):发生事件的属性值. 设事件 A 有属性 $a_1,a_2,\cdots,a_k$,则 $a_i(1\leqslant i\leqslant k)$ 为项.

定义 2　项集(item Set):若干个项(item)组成的非空集合,表示为 $I=\{a_1,a_2,\cdots,a_k\}$,其中 $a_i(1\leqslant i\leqslant k)$ 是项集中的项,也称为项集中的元素(element).

定义 3　序列(sequence):不同项集(item Set)的有序排列,表示为 $S=\langle I_1,I_2,\cdots,I_m\rangle$,其中 $I_i(i=1,2,\cdots,m)$ 为非空项集,也称为序列中的 1 个元素. 序列的长度指 1 个序列中包含的所有元素的个数,

收稿日期:2012-11-16
基金项目:教育部重点实验室基金(110411),广东省自然科学基金(10451009001004804)和广东省科技计划(2012B091000173)资助项目.
作者简介:滕少华(1962-),男,江西南昌人,教授,博士,主要从事协同计算、数据挖掘和网络安全方面的研究.

表示为 $l = \sum_{i=1}^{m} I_i$. 长度为 k 的序列称为 k-序列. 项集可以看做是长度为 1 的序列.

定义 4 序列数据集(sequence data Set):序列数据集 D 是元组 $\langle S_{id}, S\rangle$ 的集合,其中 S_{id} 是对应序列的序列号. 序列数据集中元组的个数称为该序列数据集的大小,记为 L.

定义 5 支持度(sup):一个序列 S 在序列数据集 D 中的绝对支持度指 D 中包含 S 的元组数目,记为 $sup(S)$;相对支持度指 D 中包含 S 的元组在整个数据集元组中所占的百分比,即 $sup(S)/L$. 本文若不特别申明,均指绝对支持度.

定义 6 序列模式(sequence pattern):对于一个序列 S,给定一个最小支持度阈值 min_sup,如果 $sup(S) \geqslant min\_sup$,则序列 S 是序列模式.

此处本文假设序列数据集为 D,最小支持度为 min_sup. 序列模式挖掘的任务就是要找出 S 中所有的序列模式.

1.2 序列模式挖掘算法分析

序列模式挖掘算法大体可以分为 2 类: 一类是基于 Apriori 类的,主要是应用 Apriori 类算法的穷举原则和特性,如 AprioriALL 算法[6];另一类算法则是基于模式增长的原理,运用分而治之策略,重复将数据库投影至更小的数据集中,然后在此较小数据集上应用扩展序列模式挖掘,如 PrefixSpan 算法、Freespan 算法等[7-12]. 目前效率较高且广泛应用的序列模式挖掘算法为 PrefixSpan 算法,该算法依靠各频繁前缀子序列的投影,通过递归过程逐步缩小投影序列集的规模,从而大大提高了生成序列模式的效率. 因此,本文采用 PrefixSpan 算法进行实验.

2 警用车辆维修系统的序列模式挖掘

2.1 警用车辆维修系统简述

警用车辆维修系统是用于管理警用车辆维修申请与审批流程. 各下级大队通过上传车辆维修报价单和车辆维修部件明细向各层上级提出维修申请,各上级部门逐层根据业务需求给予审核与批示,其系统业务流程如图 1 所示.

在此业务流程中会产生大量各基层单位车辆维修记录,其中包括车辆每次维修价格及每次维修所更换的汽配零部件. 这些维修记录就作为本次研究与分析的数据源.

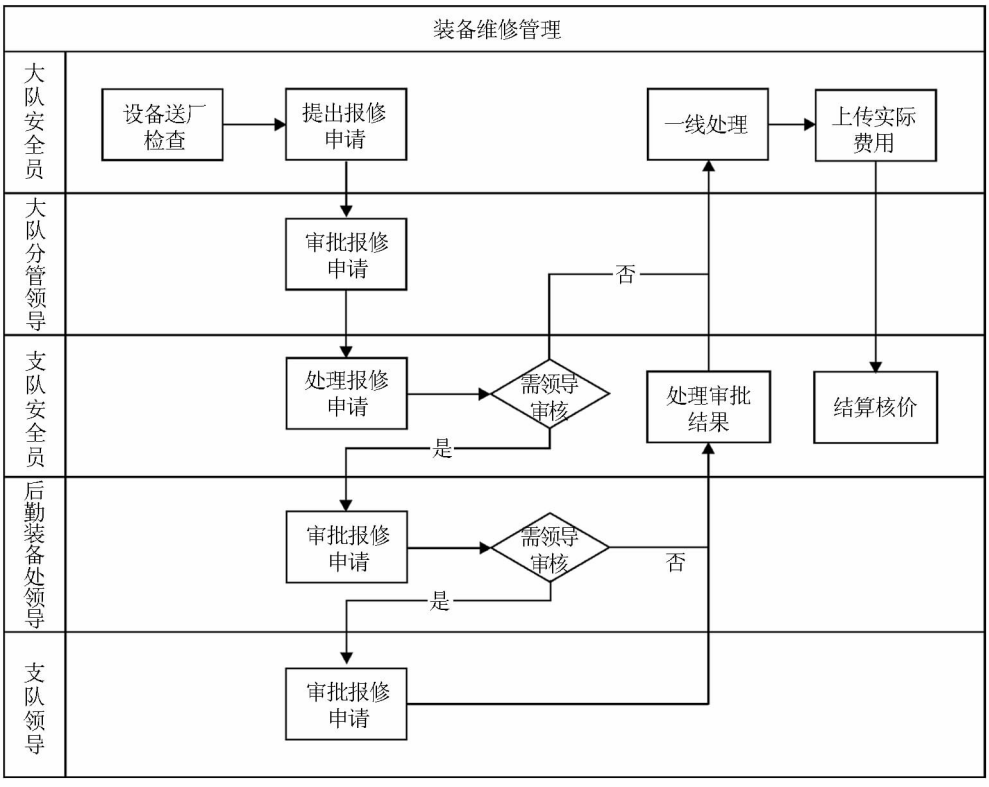

图 1 警用车辆维修管理业务流程图

2.2　警用车辆维修系统的挖掘模型

警用车辆维修序列的挖掘过程主要分为以下几个步骤: 数据收集、数据预处理、数据挖掘和存入规则库 4 个步骤,其体系结构图如图 2.

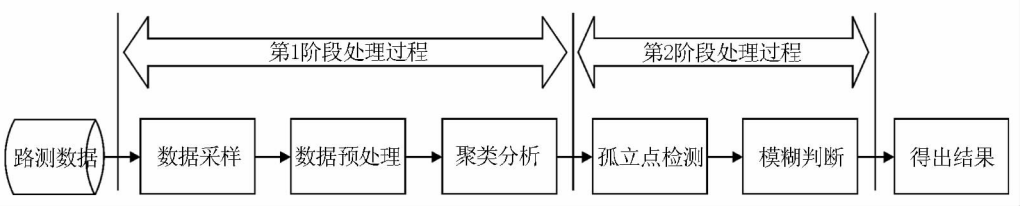

图 2　挖掘模型图

2.3　数据提取

在警用车辆维修系统中,各申请报修的车辆需要上传一个维修报价单,该报价单中包含的信息有: 车牌号、车辆识别码、维修汽配件明细,以及相应的价格开销等相关信息. 这些信息存放在数据库的相应表中.

然而,本次研究与分析的对象是车辆每次维修所更换零配件序列模式及车辆每次维修价格序列模式. 所以将与研究相关的表和属性作连接操作,提取出相关属性,各字段解释如表 1 所示.

表 1　数据提取处理之后的属性情况

属性名称	属性描述	属性类型
sReportID	车辆报修事务编号(用于表示每一次报修事务)	discrete
sDeviceReportID	车辆的车牌号码	discrete
sReportPerson	车辆识别码(和车牌号码一起标识唯一一辆车)	discrete
dtReportTime	报修事务的申请时间	date
fTotalCost	每次维修事务的花费	continuous
sAccessoryName	每次维修事务所更换的汽配零件	discrete

2.4　数据预处理

由于各种主、客观或不可抗拒等原因,数据提取过程获取的数据往往存在不一致、冗余、不完整的现象,直接对这些数据进行序列模式挖掘几乎不可能. 需要通过数据预处理操作对采集到的数据进行数据清理、数据采样等一系列操作,使之满足序列模式挖掘要求. 通过数据预处理改进了数据的质量,当然也可能损失部分数据,但有助于提高后期序列模式挖掘的精度和性能.

本文提取的数据表中汽配件的表述不够规范,需要把这些数据对应到规范的名称上,方法如下:

(ⅰ) 预先设定好一个汽配件集合 S

(ⅱ) for each$A_n \in$ sAccessory Name 列的所有数据行

(ⅲ) for each $s \in S$

(ⅳ) 找出 S 中与 A_n 匹配度最高的;

　　//匹配度根据字符串比较字数的大小

(ⅴ) 当前行的 sAccessory Name 设置为当前汽配件变量 s.

对于警用车辆维修费用属性数据的序列模式挖掘,由于其值是数值型,属于连续性数据,不适用于序列模式挖掘中项的比较,所以对于费用属性列 fTotalCost(见表 1) 的预处理方法为: 扫描提取结果中的 fTotalCost 列(见表 1) 的所有数据行,对所取数值的除 300,并将所得结果用字母对应,结果如表 2 所示.

表 2　维修费用 fTotalCost 的符号模式

数值	< 300	[300,600)	…	≥3 000
符号	a	b	…	j

2.5　序列模式挖掘

此处采用基于投影方式的序列模式挖掘算法 PrefixSpan.

(a) 输入: 车辆维修所需零配件序列或车辆维修所需费用序列 SD;

(b) 输出: 车辆维修所需零配件序列或车辆维修所需费用序列 SD 中的所有频繁模式;

(c) 实现: call procedure PrefixSpan(0,∅,SD) .

Procedure PrefixSpan(L,s,$SD|_s$)

/* s 为一个序列模式,L 为序列模式 s 的长度,$SD|_s$:如果 s 不为∅,则 $SD|_s$ 为 s 的一个投影序列数据库,否则,$SD|_s = SD$. * /

(ⅰ)扫描 $SD|_s$ 一次;

(ⅱ)if(不存在频繁项 t:将 t 加入到 s 中的最后一个元素里后所变成的序列模式是频繁的)

(ⅲ)return;

(ⅳ)for each 满足(ⅱ)的(t);

(ⅴ)(t)作为 s 中的最后一个元素后的序列模式为频繁序列模式;

(ⅵ)将加入 t 到 s 中后的频繁序列模式记为 s';

(ⅶ)对每一个 s',构造投影序列数据库 $SD|_{s'}$;

(ⅷ)for each $SD|_{s'}$;

(ⅸ)call procedure PrefixSpan($L+1$,s',$SD|_{s'}$).

3 实验与分析

本文数据源于某市警用车辆维修管理系统,该系统包括 1 263 台车辆的 5 016 次维修记录,本文分别选用了不同的支持度阀值进行实验. 并做了 2 次挖掘实验,分别是对于费用 fTotalCost 列的挖掘和 sAccessory Name 列的挖掘. 实验环境为 Intel Core Due 1.83 GHz、2.5 GB 内存,Microsoft Visual Studio 2008. 实验结果图 3 所示.

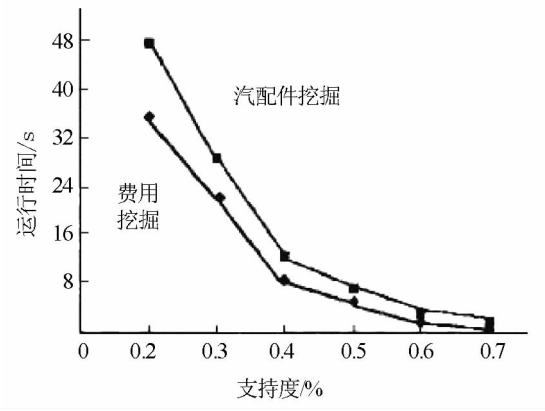

图 3 程序不同情况下的运行时间对比图

由图 3 不难得出如下实验结果:

(ⅰ)支持度和程序运行时间及产生模式成反比,即当支持度越小,程序运行的时间就越长,而且产生的模式就会越多.

(ⅱ)在支持度相同的情况下,序列长度和程序运行时间及产生模式成正比,即序列越长,程序运行时间就会越长,产生的模式就会越多.

(ⅲ)当支持度为 0.6 时,产生出的模式已经很少,当 0.7 时,生成的模式已经几乎无任何意义.

(ⅳ)当支持度阀值 0.2 时,出现爆炸式的模式,即大多数模式无意义.

(ⅴ)相同支持度阀值情况下,零配件挖掘时间比费用更长,因为维修事务中包括配件多项,而费用只包括一项.

下面给出适当支持度情况下挖掘的序列模式,如表 3 所示.

表 3 实验生成的模式

零配件 $Sup=0.5$	费用 $Sup=0.4$
〈机油,机油〉	〈a,a,d〉
〈机油,{机油格、机油}〉	〈a,b,a〉
〈机油,机油格〉	〈a,b,d〉
〈{机油、机油格},机油〉	〈a,d〉
〈{机油、机油格},{机油、机油格}〉	〈b,a〉
〈{机油、机油格},机油格〉	〈b,b〉
〈机油格,机油〉	〈b,c〉
〈机油格,{机油格、机油}〉	〈b,d〉
〈机油格,机油格〉	〈d,b,b〉
〈轮胎,机油〉	〈e,e〉
〈轮胎,{机油格、机油}〉	
〈轮胎,机油格〉	

由表 3 的实验结果可得出:对于零配件而言,机油,机油格有极高的消耗频率,机油使用以后消耗机油格,以及机油格使用完以后使用机油,和同时使用的频率是很高的. 还有轮胎消耗以后,使用机油格、机油的频率也很高.

对于维修费用而言,除了序列〈b,a〉,〈a,b,a〉以外,其它序列均表明,维修的费用呈现保持或上升的趋势.

4 总结

本文对警用车辆维修记录,提出了一种基于序列模式的挖掘方法. 实验结果表明,序列模式挖掘可用于挖掘警用车辆维修系统中的序列模式,下一步将探讨云计算下的序列模式挖掘问题.

5 参考文献

[1] Han J W,Kamber M,Pei J. 数据挖掘概念与技术 [M]. 3 版. 范明,译. 北京:机械工业出版社,2012.

[2] 李川川,刘衍珩,田大新. 基于序列模式的网络入侵检测系统 [J]. 吉林大学学报: 工学版 2007,37(1):121-125.

[3] 王令剑,滕少华. 聚类和时间序列分析在入侵检测中的应用 [J]. 计算机应用,2010,30(3):699-701.

[4] Ren J M,JSR Jang. Discovering time-constrained sequential patterns for music genre classification [J]. Audio, Speech, and Language Processing,2012,20(4):1134-1144.

[5] 俞东进,郑苏杭,李万清. 基于多核并行的海量数据序列模式挖掘 [J]. 计算机应用研究,2012,29(2):478-481.

[6] Agrawal R,Sriaknt R,Mining sequential patterns [EB/OL]. [2012-09-17]. http://rakesh.agrawal-family.com/papers/icde 95seq.pdf.

[7] 赵玉明,张巍,滕少华. 数据挖掘技术在异常检测中的应用 [J]. 微型电脑应用,2007,23(5):16-17.

[8] Pei Jian,Han Jianwei,Behzad M A,et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth [EB/OL]. [2012-09-21]. http://pdf.aminer.org/000/300/860/prefixspan_mining_sequential_patterns_by_prefix_projected_growth.pdf.

[9] Pei Jian,Han Jianwu. Mining sequential patterns by pattern-growth: the prefixSpan approach [J]. IEEE Transactions on knowledge and data engineering,2004,16(11):1424-1440.

[10] Lin Cindy Xide,Ji Ming,Danilevsky M,et al. Efficient mining of correlated sequential patterns based on null hypothesis [EB/OL]. [2012-09-21]. http://dl.acm.org/citation.cfm? id=2389656.2389660.

[11] 郭小芳,李锋,宋晓宁. 一种基于 PCA 的时间序列异常检测方法 [J]. 江西师范大学学报: 自然科学版,2012,36(3):280-283.

[12] 郭小芳,李锋,刘庆华. 一种有效的多元时间序列相似性度量算法分析 [J]. 江西师范大学学报: 自然科学版,2013,37(1):56-59.

Sequential Pattern Mining Applying in Analyzing Police Vehicle Maintenance Data

TENG Shao-hua,HONG Jia-ming,ZHANG Wei

(College of Computer,Guangdong University of Technology,Guangzhou Guangdong 510006,China)

Abstract: The sequential pattern mining method has been used to analyze the police vehicle maintenance data. The multi-dimension analysis of vehicle maintenance data has been used to extract decision support information. The experimental results show that our method is feasible and effective.

Key words: sequential pattern mining; police vehicle maintenance; data mining

(责任编辑: 冉小晓)