

智能交通系统开发中关联规则挖掘算法的应用^①

Implementation of Association Rule Mining Algorithm and Its Application in Intelligent Transportation System

肖 娟 叶 枫 (浙江工业大学 浙江 杭州 310000)

摘 要: 重点介绍了关联规则挖掘算法中的 FP-Growth(Frequent Pattern Growth, 频繁模式增长)算法, 并给出 FP-Growth 算法的编程实现以及在智能交通系统中的应用实例。研究表明, 其挖掘结果可对交通违法行为的预测提供重要的决策支持。

关键词: 关联规则 FP-growth 频繁项集 智能交通系统

随着时间的推移, 智能交通系统积累了大量原始数据, 这些海量数据中蕴藏着大量有价值的信息。数据挖掘是一种深层次的数据分析方法, 它能提取出隐藏在数据之中深层次的、关于数据整体特征描述和数据发展趋势的预测信息和知识, 把对数据的应用从低层次的联机查询操作提升到决策支持、分析预测等更高层次的应用上。数据挖掘的应用对于交警部门实时监控路面状况和交通流量, 有效疏导交通阻塞, 减少交通违法、交通事故的发生都具有重大的意义。本文对规则挖掘算法中的 FP-Growth 算法在智能交通管理系统中的应用进行了初步的探索, 获得了违法业务系统中有效的规则。

1 关联规则挖掘算法的设计与实现

1.1 FP-growth 算法

关联规则挖掘可以划分成两个子问题: 发现频繁项目集和生成关联规则。即先通过用户给定的最小支持度, 寻找满足 support 不小于 min_sup 的频繁项目集。然后通过用户给定的最小可信度, 在每个最大频繁项目集中, 寻找 confidence 不小于 min_conf 的关联规则。比较典型的关联规则挖掘算法有 Apriori 算法和 FP-growth 算法。

FP-growth 算法是 Han. JW 等人在 2000 年提出

的, 该算法特点是不生成候选集而直接生成频繁集, 将整个数据库压缩成精简的 FP 树, 且只对数据库进行 2 次扫描。相比 Apriori 算法, FP-growth 算法的效率和伸缩性都很好, 对挖掘较长的模式和短模式都有很好的适合性。由于交通管理系统中存储的数据量非常巨大, 故选用 FP-growth 方法实现关联规则挖掘较合适。

FP-growth 算法是基于这样的思想: 1、既然只有频繁项才有可能出现在频集中, 那么它通过第一次扫描数据集将所有的频繁项及其支持度找出来; 2、为避免重复扫描数据集, 在内存中维护一个相对紧凑的数据结构 fp-tree, 来存储事务记录中对应的频集。如果多个事物记录共享相同的频集, 在 fp-tree 中此频集仅被存储一次, 频集在数据集中的出现次数由相应的计数来跟踪和维护; 具有相同前缀的频集, 在 fp-tree 中也分享相同的节点, 相应节点(或者是项)的出现次数由相应的计数跟踪和维护。

本文利用 Visual C++ 为开发环境实现了 FP-Growth 算法。Visual C++ 提供了功能强大的标准模板库 STL, 且提供了更好的代码重用的机会。

尤其是其中一些算法可以删除对象, 给对象进行排序, 以及给对象记数, 进行相互比较, 然后找出特殊的对象。这些执行特性强的算法能大大减少了工作量。

^① 收稿时间: 2009-01-09

2 智能交通管理系统关联规则挖掘模型

2.1 基于概化的数据预处理

本文以智能交通信息系统为对象，应用上述算法实现对交通违法行为记录的挖掘，系统后台数据库为 Oracle 9i。涉及到的信息包括了机动车信息，车主信息，以及根据机动车车牌号码和驾驶人身份证号码从支队违法业务数据库中取得的违法记录等。考虑到数据分析的完整性，本文从系统数据库中取了 60 万条 2005 年所采集的违法纪录进行分析。原始数据来自交通信息系统中的违章信息表、机动车信息表和驾驶人信息表，表的部分字段如下：

违章信息表(车牌号，车型，违章类型，违章时间，违章地点)；

车主信息表(车主姓名，身份证号，性别，出生日期，驾驶证发放日期)；

机动车信息表(车牌号，车型，车辆品牌，身份证号码，强制报废期，车辆型号，发证日期)。

由于表中涉及到的很多字段具有连续的属性，而采用关联规则产生频繁项目集需要离散的数据，因此有必要对现有数据进行数据预处理，考虑系统的特点及用户的需求，我们采用了基于概化的数据预处理方法，从而提高挖掘效率。

对于违法日期数据，我们将月份数据概化为 M1-M12(1-12 月)，将日概化为：D1[上旬：1-10 日]，D2[中旬：11-20 日]，D3[下旬：21-31 日]。

对于时间数据，根据当地实情，将上下班车流高峰期合理地分到六个时间段：T1[早晨 5:01-9:00]、T2[上午 9:01-12:00]、T3[中午 12:01-14:00]、T4[下午 14:01-17:00]、T5[傍晚 17:01-20:00]、T6[夜间 20:01-5:00]。

系统通过车主身份证号码得到的车主年龄，按距离概化为：N1[<20 岁]，N2[20-29 岁]，N3[30-39 岁]，N4[40-49 岁]，N5[50-59 岁]，N6[≥60 岁]。

通过驾驶证发放日期按需求把驾龄概化为：L1[<1 年]，L2[1-2 年]，L3[2-3 年]，L4[3-5 年]，L5[5-10 年]，L6[10-20 年]，L7[≥20 年]。

对于其他数据采用伪代码标识，如车型、违章类型、违章地点在数据库都已有相关的表。车型表：01[大型汽车]，02[小型汽车]，03[使馆汽车]，04[领馆汽车]，05[境外汽车]，06[外籍汽车]，07[两、三轮摩托]，08[轻便摩托车]，09[使馆摩托车]，10[领馆摩

托车]，11[境外摩托车]，12[外籍摩托车]，13[农用运输车]，14[拖拉机]等 26 种类型。违章类型以 4 整数表示。违章地点以[WD1-WDn]表示，性别以 1[男]，2[女]表示。将这些数据加以整理，抽取和概化，得出表 1。

表 1 数据挖掘分析所用到的项

变量	项
违法月份	M1~M12(1 月至 12 月)
违法日期	D1,D2,D3 (上旬、中旬、下旬)
违法时间	T1[5:01-9:00]、T2[9:01-12:00]、T3[12:01-14:00]、T4[14:01-17:00]、T5[17:01-20:00]、T6[20:01-5:00]
驾驶员性别	1[男]，2[女]
驾驶员年龄	N1[<20]，N2[20-29]，N3[30-39]，N4[40-49]，N5[50-59]，N6[≥60]
驾龄	L1[<1]，L2[1-2]，L3[2-3]，L4[3-5]，L5[5-10]，L6[10-20]，(7)[≥20]
车型	01[大型汽车]，02[小型汽车]，03[使馆汽车]，04[领馆汽车]，05[境外汽车]，06[外籍汽车]，07[两、三轮摩托]，08[轻便摩托车]，09[使馆摩托车]，10[领馆摩托车]，11[境外摩托车].....26 种
违章类型	1101[未按规定使用安全带]，1303[超速 50% 以下]，1109[未随车携带行驶证]，1301[机动车逆向行驶的].....
监控点位置(违章地点)	JK1[江南工贸大街-车站路]、JK2[江南工贸大街-红旗路 1] JK3[江南工贸大街-红旗路 2]、JK4[江南工贸大街-苕溪路]、JK5[江南工贸大街的苕溪路和红旗路之间]、JK6[人民路-红旗路]、JK7[人民路-苕溪路]、JK8[人民路-益民路]、JK9[苕溪路-龙溪南路]、JK10 长桥南路在城南南路
所属单位	330502200001[粮贸城社区]、330502200200[道场滨村]、330502200201[钱山下村]、330502200202[塘南村].....
所属辖区	330502000100[直属一大队一中队]、330502000200[直属一大队二中队]、330502000300[直属一大队三中队]、330502000400[直属一大队四中队].....

预处理后的数据如表 2 所示：

表 2 预处理后的数据

车牌号	车型	违章类型	违章月份	违章日期	违章时间	违章地点	年龄	性别	驾龄
浙 EX ××10	1 4	1201	M12	D1	T2	101040 000000	N 4	1	L
浙 EX ××55	0 2	1109	M4	D2	T4	957990 006000	N 3	1	L
浙 EX ××19	0 1	1027	M6	D1	T5	957990 006000	N 3	2	L

2.2 挖掘模式及解释

数据经预处理后最终得到 6358 条有效数据, 对其进行模式挖掘, 经验设置的最小支持度 minsup 为 0.01, 最终得到 3207 项频繁项集, 最多的频繁项集包括 6 项。表 3 所示的是挖掘所得的部分最大频繁项目集。

表 3 挖掘所得的部分最大频繁项目集

序号	项	支持度
1	(1,N3,M3,D1,T2,02,1303)男 30~40 岁 3 月上旬 9 点~12 点 小型车 超速	66
2	(1,N3,M3,D1,T2,104,1303)男 3 月上旬 9 点~12 点 小型车 104 线 超速	64
3	(1,N4,M2,D3,T2,02,1303)男 40~50 岁 2 月下旬 9 点~12 点 小型车 超速	70
4	(1,N3,M3,D1,T2,104,1303)男 30~40 岁 3 月上旬 9 点~12 点 104 线 超速	66
5	(1,N3,M3,D3,T2,02,1303)男 30~40 岁 2 月下旬 9 点~12 点 小型车 超速	67

其中, 项 1303 的支持度达 32%, 说明项 1303 是比较典型的交通违法行为, 查询相应的违法代码表可以发现与项 1303 对应的是“机动车行驶速度超过规定时速 50%以下”, 由此可得超速是引起交通违法行为的主要原因。

设置最小置信度为 0.35,从最大频繁项目集中获得部分满足条件的有效关联规则如表 4 所示:

表 4 由最大频繁项目集所生成的部分关联规则

序号	规则	支持度	置信度
1	1,N3,M3,D1,T2,02->1303	1.0%	50%
2	1,N3,M3,D1,T2,10104->1303	1.0%	67%
3	1,N4,M2,D3,T2,02->1303	1.1%	58%
4	N3,T2->1303	6.1%	35%
5	T2->1303	14.1%	35%
6	1,N3,M3,D3,T2,02->1303	1.0%	52%
7	1,D3,02,95799->1303	1.1%	69%
8	95799->1303	11.7%	40%
9	1,N3,M2,65299->1101	1.1%	36%
10	1,M3,02,65299->1101	1.0%	36%
11	1,M1,02,65299->1101	1.0%	38%
12	N3,02,10104->1071	1.0%	35%
13	1,M1,02,10104->1071	1.0%	35%
14	1,N4,M2,65299->1101	1.3%	37%

分析挖掘的结果我们可以得到如下规则:

R1:1,N3,M3,D1,T2,02->1303,表明年龄在 30~40 岁的男性驾驶员在 2005 年 3 月上旬早上 9:00~12:00 之间行驶小汽车超速的可能性为 50%, 支持度为 1.0%。

R2:1,N3,M3,D1,T2,10104->1303,表明年龄在 30~40 岁的男性驾驶员在 2005 年 3 月上旬早上 9:00~12:00 之间行驶在 104 国道上超速的可能性为 67%, 支持度为 1.0%。

R3:1,N4,M2,D3,T2,02->1303,表明年龄在 40~50 岁的男性驾驶员在 2005 年 2 月下旬早上 9:00~12:00 之间行驶小汽车超速的可能性为 58%, 支持度为 1%。

R4:N3,T2->1303,表明年龄在 30~40 岁的驾驶员在早上 9:00~12:00 之间超速的可能性为 35%, 支持度为 6.1%。

R5:T2->1303,表明驾驶员在早上 9:00~12:00 之间超速的可能性为 35%, 支持度为 14.1%。

.....

R7:1,D3,02,95799->1303,表明男性驾驶员在 2005 年每月的下旬期间驾驶小型汽车在城郊超速的可能性为 69%, 支持度为 1.1%。

R8:95799->1303,表明驾驶员在城郊超速的可能性为 40%, 支持度为 11.7%。

.....

R13:1,M1,02,10104->1071,表明男性驾驶员在 2005 年 1 月期间驾驶小型汽车在 104 国道上因未按规定使用安全带而违反交通法规的可能性为 35%, 支持度为 1.0%。

.....

2.3 模式评价

由上述挖掘出的模式可以看出:

① 在地点 95799 超速的可能性比较大。实际上该地点位于城郊, 离市中心较远, 道路比较宽阔, 很少有建筑物和行人、交通管治相对比较松懈,所以车辆超速比较频繁。

② 违章记录中, 驾驶员年龄在 30~45 岁早晨之间车辆超速现象比较严重, 进一步分析发现这段时间是早上上班高峰期,且大多数路段交警的管理力度都还跟不上, 因此可在这一时段加强警力。

3 结论

本文将关联规则挖掘算法运用于交通违法系统的模式挖掘, 并进行了初步地探索和研究, 结果表明, 我们所挖掘的模式是有效的, 我们可将挖掘结果用于智能交通系统中, 为交通管理提供决策支持具有重要的实用价值。

参考文献

- 1 尚威, 尚宁, 覃明贵. 交通事故的多维关联规则分析. 计算机应用与软件, 2006, 23(2): 40 - 42, 65.
- 2 李英, 李武, 王浣尘. 聚类分析算法在交通控制中的应用. 系统工程, 2004, 22(2): 66 - 68.
- 3 谢嘉孟, 彭宏, 周兵, 杨东辉. 基于数据挖掘技术的智能交通信息分析与决策研究. 公路, 2004(4): 154 - 158.
- 4 刘淑霞, 陈燕. 数据挖掘在交通肇事逃逸案中的应用. 刑事技术, 2005(2).
- 5 Chang LY, Chen WC. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 2005(36): 365 - 375.
- 6 Chong MM, Abraham A, Paprzycki M. Traffic accident analyses using decision trees and neural networks.
- 7 Lee DH, Jeng ST. Applying data mining techniques for traffic incident analysis. Journal of the Institution of Engineers, 2004.
- 8 Lu J, Cao L. Congestion evaluation from traffic flow information based on fuzzy logic. IEEE, 2003.
- 9 Geurts K, Thomas I, Wets G. Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention, 2005, (37): 787 - 799.
- 10 Li TR. Mining ϕ -frequent itemset using FP-tree. Journal of Southwest Jiaotong University(English Edition), 2001, 9(1): 67 - 74.
- 11 Yang M, Yang P, et al. Fast FP-Growth for association rule mining. Journal of Southeast University