

数据挖掘技术在二手车交易系统中的应用

陈 君

(渭南师范学院 网络安全与信息化学院 网络安全与信息化工程技术中心,陕西 渭南 714000)

摘 要:数据挖掘技术是指从数据集中发现有效的、新颖的、潜在有用的和最终可以理解模式的高级处理过程,FP-growth 算法是数据挖掘算法的一种。FP-growth 算法是一种基于 FP-tree 的频繁项集挖掘算法,此算法是将原始数据集压缩到一棵 FP-tree 上,对原始数据集进行两次扫描,挖掘过程不产生候选项集,不用候选测试的算法,它使用紧缩的数据结构,避免了对数据库的重复扫描,运算速度快。文中收集了乐购二手车交易平台 2016 年 1 月到 2018 年 12 月共 3 年的数据,系统中可供挖掘的模块包括:二手车信息模块,拍卖品管理模块,购物车管理模块,订单管理等信息模块。利用 FP-growth 算法对乐购二手车交易系统数据库中的车辆品牌、使用年限、车载人数、行驶里程、车辆价格、保养状况等信息进行整理、转换、对比、分析,从中发现二手车交易中的规律,挖掘用户购车和卖车的有关规律,提高了车辆的成交率。

关键词:FP-growth 算法;二手车交易系统;关联规则;数据挖掘;FP-tree

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2020)05-0180-05

doi:10.3969/j.issn.1673-629X.2020.05.034

Application of Data Mining Technology in Second-hand Car Trading System

CHEN Jun

(Network Security and Information Engineering Technology Center, School of Network Security and Informatization, Weinan Normal University, Weinan 714000, China)

Abstract: Data mining technology is an advanced process that discovers effective, novel, potentially useful, and ultimately understandable patterns from data sets. As one of the data mining algorithms, FP-growth algorithm is a frequent item set mining algorithm based on FP-tree, which compresses the original data set onto a FP-tree and scans the original data set twice. The mining process neither generates a candidate set nor uses the candidate test algorithm. It uses a compact data structure, which avoids repeated scans of the database and has fast operation speed. We collect data from Tesco's used car trading platform for three years from January 2016 to December 2018. The modules available for mining in the system include used car information module, auction management module, shopping cart management module, order management and other information modules. We adopt FP-growth algorithm to sort, convert, compare and analyze the information of the vehicle brand, service life, number of people on board, driving mileage, vehicle price and maintenance status in Tesco Used Vehicle Trading System database, from which we can find the rules of second-hand car transactions and excavate the relevant rules of buying and selling cars, to improve the turnover rate of vehicles.

Key words: FP-growth algorithm; second-hand car trading system; association rule; data mining; FP-tree

0 引言

可以通过二手车交易系统进行二手车购买、二手车出售、二手车拍卖、搜索二手车信息、了解二手车资讯、讨论二手车问题,通过使用二手车交易系统方便进行二手车的买卖。那么哪些品牌、哪些价位、具体有哪些性能指标的二手车的成交率比较高,哪些二手车辆

更适合自己,文中针对乐购二手车交易系统数据库中的数据进行了分析挖掘,找出乐购二手车交易系统的有效规律,提高乐购二手车交易网的成交率。

利用 FP-growth 算法,对乐购二手车交易系统数据库中的车辆品牌、车载人数、行驶里程、使用年限、车辆价格等信息进行挖掘,从中发现乐购二手车交易系统数据库

收稿日期:2019-06-06

修回日期:2019-10-09

网络出版时间:2020-01-10

基金项目:渭南师范学院自然科学类研究项目(17YKP09);陕西省自然科学基金面上项目(2017JC2-08);国家自然科学基金面上项目(61572392)

作者简介:陈 君(1982-),女,讲师,硕士,CCF 会员(92336M),研究方向为数据挖掘。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20200110.1121.024.html>

有效规律,提高车辆的成交量^[1]。例如:哪些年龄段的人喜欢购买什么品牌的车,哪些收入水平的人喜欢购买什么类型的车,哪些车型的车受到男/女性的青睐、什么颜色的车辆更加容易交易、什么车龄的车在二手车交易市场比较好卖、车况的保养情况对车辆交易的影响等等。经过对相关数据的分析有助于购买方和出售方进行车辆交易,提高乐购二手车交易系统的成交率。

1 数据挖掘及关联规则的概念

数据挖掘(data mining)是指从数据集中发现有效的、新颖的、潜在有用的和最终可以理解模式的高级处理过程^[2]。数据挖掘也可以理解为从大量的数据中提取或者挖掘知识,而在数据挖掘中所提取的有价值的信息或者知识,除了一般所讲的数据和信息以外,还有广泛意义上的概念、模式、规律、约束、规则等内容^[3]。数据挖掘技术可以帮助用户进行决策、查询处理、信息管理和过程控制等。数据挖掘技术已经在市场销售、科学应用、欺诈甄别、Internet 的应用、金融等领域发挥了作用,并将成为以下行业(如网络服务、商业智能、生物工程)的研究方向。数据挖掘技术是信息技术产业最有前途的交叉学科之一。

数据挖掘过程可大体分为以下几个步骤^[4],如图 1 所示。

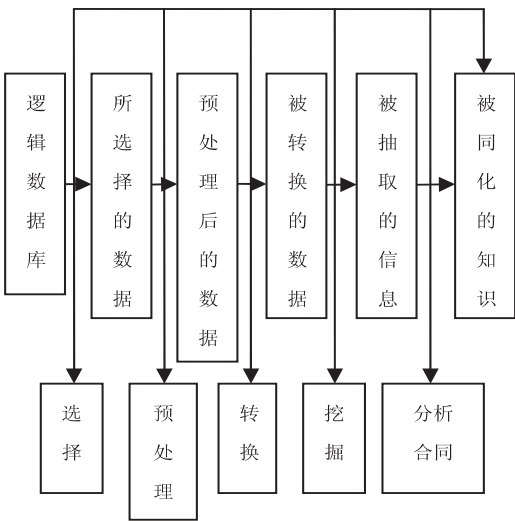


图 1 数据挖掘的步骤

- (1)业务对象:对业务问题做了清晰的定义,数据挖掘中最为重要的一步就是了解数据挖掘的主要目的,进行数据挖掘的过程中结果是不可预测的,但问题是可预知的。
- (2)数据准备:数据的准备关系到数据挖掘的结果成功与否,会不会产生经济效益。数据的准备是从不同的数据源中整理数据挖掘过程所需的数据,保证数据的易用性、时效性、综合性和数据的质量。数据挖掘的经验和工具决定了数据挖掘的结果,而数据的准

- 备十分重要。
- 数据准备可分为以下几部分:
 - ①数据的选择:查找并收集和业务对象相关的数据信息,筛选有效的数据使其能够进行有效的数据挖掘。
 - ②数据的预处理:对选择出有质量的数据进行分析研究,确定数据的类型,为下一步的数据操作做好准备。
 - ③数据的转换:创建适合具体挖掘算法的分析模型是数据挖掘结果成功与否的关键所在,在数据挖掘过程中针对数据挖掘算法创建了分析模型数据,方便用户将数据转换成分析模型。
 - (3)数据挖掘:选择正确的数据挖掘算法对转换后的数据进行挖掘取得结果。
 - (4)结果分析:对数据挖掘后的结果进行解释并做出评估,而使用的分析方法将根据数据挖掘的操作来确定,一般会用到可视化的操作技术。
 - (5)知识的同化:把分析研究所得到的知识放到业务信息系统中来考量。

关联规则属于数据挖掘之中的一种,关联规则的有效性规则为用户设定合适的支持度和置信度,产生大于最小取值的有效性规则^[5]。下面来举例说明一下什么是关联规则,比如设定牛奶和面包的支持度 support 为 15%,置信度 confidence 为 75%,也就是说来超市的顾客中 15% 的顾客同时购买了牛奶和面包,而其中购买牛奶的顾客中有 75% 同时购买了面包^[6]。

关联规则挖掘的步骤:①发现所有频繁项集。首先设定最小支持度,找出所有频繁项集,找出 support 大于等于最小支持度的所有的项目子集。实际中频繁项集之间会存在包含的关系,通常情况下只关心不被其他频繁项集所包含的那些最大频繁项集的集合,在数据挖掘中发现所有的频繁项集是关联规则形成的基础。②生成关联规则。首先用户给定一个合适的最小置信度,然后在最大频繁项集的基础上,找出 confidence 大于等于最小置信度的关联规则。

关联规则挖掘的基本模型如图 2 所示。

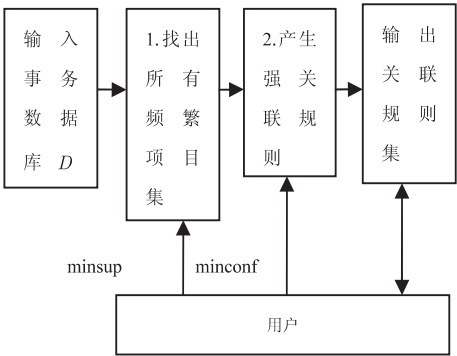


图 2 关联规则挖掘的基本模型

输入事务数据库 D , 首先根据选定算法找出频繁项目集, 生成强关联规则, 输出关联规则集合。根据用户指定的最小支持度 $\min_support$ 和最小置信度 $\min_confidence$ 分别与找出所有频繁项目集和产生强关联规则进行交互, 然后通过与输出关联规则集的交互对所得到的结果做出解释与评估。挖掘关联规则的关键步骤是步骤 1 找出所有频繁项目集, 步骤 1 的性能决定了关联规则挖掘的整体性能。所以目前的研究一般都集中在对频繁项集的挖掘和处理上面, 对比步骤 1, 步骤 2 相对容易实现些, 步骤 2 只需从已挖掘出的频繁项集中, 举出所有可能的关联规则, 最后根据用户设定的最小支持度阈值 $\min_support$ 与最小置信度阈值 $\min_confidence$ 来考量以上关联规则, 得出有效规则^[7]。

2 FP-growth 算法

2.1 算法基本思想

韩家炜等人在 Apriori 算法的基础上提出了 FP-growth 算法, 它是一种基于 FP 树的频繁项目集挖掘算法, 首先把原数据集压缩到一棵 FP 树上, 然后对原始数据集进行二次扫描, 而数据挖掘的整个过程不产生候选项目集, 所以此算法大大提高了数据挖掘的效率^[8]。把发掘长模式的问题转换变成递归的发掘短模式的问题, 连接最不频繁的项作为后缀, 来确定最好的选择, 减少了搜索的开销^[9]。FP-growth 算法跟 Apriori 算法比起来效率提高了许多, 因为 FP-growth 算法不会产生候选集, 省去了候选测试, 数据结构相对紧缩, 不用对数据库进行重复扫描^[10]。

2.2 算法描述

FP-growth 算法分为 2 个阶段: 第一阶段构造 FP-tree, 第二阶段挖掘 FP-tree。

构造 FP-tree 的方法分两步: 第一步, 扫描数据库 D , 得出结果集 L 。第二步, 创建根节点 $null$, 选择事务 Trans 中的频繁项, 对结果集 L 进行排序, 设排序后的频繁项列表为 $[\rho \mid P]$, 支持度计数最大的即第 1 个元素为 ρ , 剩余元素的表为 P , $insert_tree([\rho \mid P, T])$ 的调用方法是: 假设 T 有子女 N 而 $N.item_name = \rho.item_name$, 则 N 的计数+1, 如果 \neq 创建新节点, N 的计数=1, 将其链接到父节点 T 上, 通过节点链结构把它链接到有相同节点的 $item_name$ 上, 假如 $\rho \neq null$, 递归调用 $insert_tree(P, N)$ ^[11-12]。

第二阶段为挖掘 FP-tree, 调用 FP-growth (FP-tree, $null$):

Procedure FP-growth (Tree, α)

①if Tree 有单个路径 p then

②for 路径 p 中节点的每个组合为 β ;

③产生模式 $\alpha \cup \beta$, 支持度 $support = \beta$ 中节点的最小支持度;

④else for each α 在 Tree 的头部;

⑤产生一个模式 $\beta = \alpha \cup \alpha$, 支持度 $support = \alpha.support$;

⑥构造 β 的条件模式基, 再构造 β 的条件 FP-Tree Tree β ;

⑦if Tree $\beta \neq \emptyset$ then

⑧调用 FP-growth (Tree β, β); }

构造长度为 1 的频繁模式, 并从它开始构造条件模式基和条件 FP-tree, 并且在此树上进行递归挖掘, 通过后缀模式和条件 FP-tree 产生的频繁模式进行模式增长的连接^[13]。FP-growth 算法不会产生候选集, 省去了候选测试, 数据结构相对紧缩, 不用对数据库进行重复的扫描, 降低了搜索开销, 提高了挖掘效率^[14-15]。

3 FP-Growth 算法的应用

文中采用 Windows 10 的操作系统, 在 Microsoft Visual Studio 2015 开发平台中使用 C#语言, 在计算机 CPU 为 intel 2.6 GHz, 内存为 4 GB 的基础上, 应用 FP-growth 算法对乐购二手车交易系统数据库中的数据进行数据挖掘得出有用的结论。

3.1 数据准备

文中收集了乐购二手车交易平台 2016 年 1 月到 2018 年 12 月共 3 年的数据。乐购二手车交易平台的数据中可供挖掘的模块有: 二手车信息模块、拍卖品管理模块、购物车管理模块、订单管理等信息模块。

3.2 数据的预处理

数据库中未经处理过的原始数据虽包含研究所需内容, 但仍然存在不足之处, 如在乐购二手车信息模块中包含车辆品牌、车辆类型、行驶里程、车辆颜色、车辆价格、车载人数、使用年限、保养状况、出售人姓名等信息, 但对于数据挖掘来说出售人姓名是没有挖掘价值的; 在购车信息模块中同样对于数据挖掘来说购车人姓名、地址等信息是没有价值的。各模块中都包含了一些无用的信息, 这些信息将会严重影响数据挖掘的效率, 因此应先进行数据的预处理操作。

3.2.1 删除无效数据

删除无效数据的操作如下:

(1) 删除表中无用的数据属性, 例如乐购二手车信息中的车辆出售人姓名, 购车信息中的购车人姓名和地址等对于本项目挖掘目标意义并不大, 可以忽略的这些字段。

(2) 删除各表中的无用数据、脏数据、不完整和不一致的数据, 例如乐购二手车交易系统中用户的注册

信息不完整、错误、前后不一致的数据。

3.2.2 数据整理和转换

对乐购二手车交易系统的数据进行整理和转换。首先整理了乐购二手车交易信息中的车辆品牌、车辆价格、行驶里程、车辆类型、车辆颜色、车载人数、使用年限、保养状况、购车人性别、购车人年龄、购车人职业等。

关联规则挖掘算法要求数据应为布尔型,而原始数据表中的数据不是,为了能够使用上述的关联规则算法对该数据表进行挖掘,需要对乐购二手车交易系统 中的原始数据进行转化:

(1)量化属性离散化:关联规则要求将数据库中的一部分数值型的属性区间化。例如将根据取值的分布规律,将数值型的属性行驶里程离散化,将它划分为 9 组分别为:30(4 万公里以下)、31(4 万到 8 万公里以下)、32(8 万到 12 万公里以下)、33(12 万到 20 万公里以下)、34(20 万到 30 万公里以下)、35(30 万到 40 万公里以下)、36(30 万到 40 万公里以下)和 37(40 万到 50 万公里以下)和 38(50 万公里以上)。其他的数值属性也按本办法,把数值属性转化为布尔型,划分成几个区间,转换成数字。

(2)类别属性转化:一些备选项属性是需要进行类别转换的,比如性别属性,把它们转化成布尔类型数据,例如:58(男)、59(女)。将数据库中其他类似的类别属性也转化成布尔型数据。

下面将举例进行说明,记录的字段名含义如表 1 所示,对应关系如表 2 所示,数据转换后的事务数据如表 3 所示。

表 1 记录的字段名含义

| 字段名 | 含义 | 内容 |
|----------------|------|-------|
| car_brand | 车辆品牌 | 本田 |
| car_type | 车辆类型 | SUV |
| car_kilometers | 行驶里程 | 4 万公里 |
| car_colour | 车辆颜色 | 白色 |
| car_price | 车辆价格 | 8 万 |
| ... | ... | ... |

表 2 对应关系

| ITEM | 属性名称 | 项目代码 |
|------|------|------|
| | 车辆品牌 | |
| 1 | 奔驰 | 1 |
| 2 | 本田 | 2 |
| 3 | 大众 | 3 |
| 4 | 奥迪 | 4 |
| 5 | 福特 | 5 |
| ... | ... | ... |

表 3 转换后的事务数据

| D | Item |
|-----|---------------------------|
| 1 | 2,15,21,34,43,73,62,78,96 |
| 2 | 2,15,21,34,43,63,47,78,96 |
| 3 | 2,15,21,34,66,48,58,75,96 |
| 4 | 2,15,21,34,66,48,58,75,96 |
| 5 | 2,15,21,34,68,48,59,75,96 |
| ... | ... |

3.3 结果及分析

对整理和转换后的乐购二手车交易系统的数据进行有效性挖掘,输入事务数据,设定最小支持度 $s = 8\%$,设定最小置信度 $c = 25\%$,输出频繁项集。挖掘结果如表 4 所示。

表 4 挖掘结果

| 规则 | 车辆品牌 | 车辆类型 | 车辆颜色 | 行车里程 | 行车年限 | 性别 | 支持度 | 置信度 |
|-----|----------|------|-----------|------------|------------|----------|-----|-----|
| A | I2 本田 | SUV | I15 白色 | I31 4-8 | | | 10 | 48 |
| B | I3 大众 | | | | I45 2-5 | I58 男 | 8 | 25 |
| C | I4 奥迪 | 轿车 | I16 黑色 | | | | 9 | 27 |
| D | I1 奔驰 | 轿车 | I17 银色 | | | I59 女 | 36 | |
| E | | | I16 黑色 | I31 4-8 | I45 2-5 | | 12 | |
| F | I4 别克 | MPV | I19 蓝色 | | | | 49 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

根据表 4 可以得出以下结论:规则 A 表示白色本田 SUV 行驶里程在 4 万到 8 万公里的二手车辆比较

受到市场的欢迎,规则 B 表示行车年限在 2 年到 5 年的大众车辆比较受男士的亲睐,规则 C 表示黑色奥迪轿车受到购车者的青睐,规则 D 表示银色奔驰轿车比较受女士的亲睐,规则 E 表示年限在 2 年到 5 年的黑色车辆,并且行车里程在 4 万到 6 万公里的比较受购车者的欢迎,规则 F 表示蓝色别克 MPV 型车辆在二手车市场上交易数量较多。

4 结束语

利用 FP-growth 算法对乐购二手车交易系统中的数据进行了对比和分析,发掘其中的有效规律,为购买和出售二手车辆的双方提供了有用的信息。在使用 FP-growth 算法对乐购二手车交易系统事务数据库进行数据挖掘的过程中,数据的准备和选择是极为关键的一步,设定合适的最小支持度和最小置信度,会直接影响到数据挖掘的结果是否有效,如果最小置信度和最小支持度的数值设定过小,数据挖掘的结果就会出现大量无效的规则,浪费资源、影响效果,如果最小支持度和最小置信度的数值设定过大,将找不出相关联的有效规则,达不到数据挖掘的最终目的。

参考文献:

- [1] 韩家炜, KAMBER M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001:160-161.
- [2] 陶雪娇,胡晓峰,刘 洋. 大数据研究综述[J]. 系统仿真学报,2013,25(S1):142-146.
- [3] 刘 莹. 基于数据挖掘的商品销售预测分析[J]. 科技通报,2014(7):140-143.
- [4] 李 聪. 物流信息大数据分析研究方法研究及应用[D]. 武汉:武汉理工大学,2014.

(上接第 179 页)

- [6] 陈凌俊. 基于物联网的居家协同智能防盗系统[J]. 电子技术与软件工程,2018(9):28.
- [7] SENNOU A S, BERRADA A, SALIHALJ Y, et al. An interactive RFID-based bracelet for airport luggage tracking system[C]//2013 4th international conference on intelligent systems, modelling and simulation. Bangkok, Thailand: IEEE,2013:40-44.
- [8] VASTIANOS G E. An RFID-based luggage and passenger tracking system for airport security control applications[C]//SPIE defense + security. Bellingham: [s. n.], 2014: 901-914.
- [9] ZOLTAN K. Using RFID and GIS technologies for advanced luggage tracking[J]. SEA-Practical Application of Science, 2015,2(8):229-234.
- [10] MERCADO F C. Pack and track;U. S,9907377 B2[P]. 2018-

- [5] 李敏杰. 基于大数据下的寄递物流管理信息系统的研究[D]. 南京:南京邮电大学,2014.
- [6] 姚丹丹. 基于数据挖掘的红塔集团数据库营销系统的研究与实现[D]. 杭州:浙江理工大学,2014.
- [7] 何柏英. 云计算环境下物流路径数据挖掘研究[D]. 合肥:合肥工业大学,2013.
- [8] 胡 森. 基于 Hadoop 的物流车辆运输监控数据管理研究[D]. 大连:大连海事大学,2014.
- [9] 毛国君. 数据挖掘原理与算法[M]. 北京:清华大学出版社,2005:42-48.
- [10] 陈兴蜀,张 帅,童 浩,等. 基于布尔矩阵和 MapReduce 的 FP-Growth 算法[J]. 华南理工大学学报:自然科学版,2014,42(1):135-141.
- [11] YANG Guangming, FENG Xiao, YANG Kun. Hydraulic metal structure health diagnosis based on data mining technology[J]. Water Science and Engineering,2015,8(2):158-163.
- [12] HAO Ming, CHENG Tiejun, WANG Yanli, et al. Web search and data mining of natural products and their bioactivities in PubChem[J]. Science China(Chemistry), 2013, 56(10): 1424-1435.
- [13] WU Qiang, XU Hua. Three-dimensional geological modeling and its application in Digital Mine[J]. Science China(Earth Sciences), 2014, 57(3):491-502.
- [14] TAHAT A, MARTI J, KHWALDEH A, et al. Pattern recognition and data mining software based on artificial neural networks applied to proton transfer in aqueous environments[J]. Chinese Physics B, 2014, 23(4):414-425.
- [15] WANG Jing, ZHAO Shenghui, XIE Xiang, et al. Mapping methods for output-based objective speech quality assessment using data mining[J]. Journal of Central South University, 2014, 21(5):1919-1926.

03-06.

- [11] GHAZAL M, ALI S, HANEEFA F, et al. Towards smart wearable real-time airport luggage tracking[C]//International conference on industrial informatics and computer systems. Sharjah: IEEE, 2016:1-6.
- [12] 龚江涛,孙世华,汤芸睿,等. 智能行李箱:中国,204181134 U[P]. 2015-03-04.
- [13] 丁世豪,李光顺,刘鹏坤,等. 基于蓝牙 4.0 的自动跟踪智能行李箱设计[J]. 电子技术,2018,47(5):47-49.
- [14] 赵艳妮,马 顺,张书源,等. 基于视觉传感器的自动跟随行李箱设计[J]. 智能城市,2017,3(8):23-24.
- [15] MADGWICK S O H, HARRISON A J L, VAIDYANATHAN A. Estimation of IMU and MARG orientation using a gradient descent algorithm[C]//IEEE international conference on rehabilitation robotics. Zurich: IEEE, 2011:1-7.