改进关联规则算法在烟草物流销售规律中的应用①

钱慎一, 王欢欢, 杨铁松

(郑州轻工业学院 计算机与通信工程学院, 郑州 450002)

摘 要: 烟草产品在销售过程中产生了大量的历史数据, 是否能挖掘出其中潜在的高价值信息对烟草物流公司至关重要. 本文采用一种比较高效的 fp-growth 算法设计了烟草产品销售决策支持系统, 实践应用表明改进后的 fp-growth 算法能快速的发现不同时期烟草产品之间的销售规律和关系, 占用内存更小、响应速度更快, 为烟草物流公司及时掌握烟草产品之间的销售规律提供了有价值的参考.

关键词: 关联规则算法; fp-growth 算法; 烟草; 物流; 应用

Application of Improved Association Rules Algorithms to Tobacco Sales Logistics in the Law

QIAN Shen-Yi, WANG Huan-Huan, YANG Tie-Song

(School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: It is vital to tobacco logistical company whether you can dig the potential high valued information from the substantial historical data that create during the tobacco products selling. This article adopts a more efficient algorithm that of fp-growth to design sales of tobacco products decision and support system, and the practice indicates the improved fp-growth algorithm is able to discover the sales discipline and relationship of tobacco products in various terms. In the algorithm, memory occupied is lower and respond speed is faster, which provides valuable references for tobacco logistical company on mastering sales discipline of tobacco products timely.

Key words: algorithm of association rules; Fp-growth; tobacco; logistics; application

随着物流企业信息化建设和应用的普遍推广, 日常物流服务中产生并积累了大量的数据和信息, 由于物流企业普遍缺乏对于业务数据分析和模式挖掘的战略重视[1], 再加上传统分析的局限性, 使得物流企业难以挖掘客户关系中潜在的有效信息, 并做出正确决策, 严重影响和制约了物流行业的发展. 目前, 国家实行烟草专卖制度, 每一地区由烟草物流公司进行统一配送, 然后由零售商进行销售. 要想使烟草物流公司利润最大化, 就要根据市场的变化, 进行更合理的采购和配送.

我国在物流业中的数据挖掘技术的研究始于 90 年代中期, 到 90 年代后期, 逐步形成了基本框架. 从 90 年代中后期, 逐渐发表在《计算机学报》、《人工智能与模式识别》等刊物上的一批研究成果, 使研究重点从以前的发现方法开始偏向于系统应用; 开始注重多种发现策略、技术集成、各种学科间的互相渗透; 但是实际应用还处于起步阶段, 主要以学术研究为主. 在国内, 适应现代物流市场的内外部环境基本还处于起步发展的阶段, 我国物流企业在数据挖掘应用方面还处于起步阶段, 经验不足, 应用实践在国内物流企业中并不多见, 当前数据挖掘在物流企业中的应用主要集中在市场预测, 物流中心的选择, 优化配送路线、合理安排商品的仓储、顾客价值分析、物流需求几方面[2].

在许多 Apriori 算法的改进算法里, FP-Growth 算法是一种无候选集的方法, 该算法是韩等人在 2000 年提出的[3], 在对 Apriori 算法及一些改进算法进行研究后, 韩等人发现 Apriori 及其改进算法的性能和采用候

① 基金项目:河南省科技攻关项目(142102210156,122102210024)
收稿时间:2015-07-01;收到修改稿时间:2015-09-08

选集有很大关系, 因此提出了基于频繁模式树 FP-Tree 的无候选集的 FP-Growth 算法, 尤其对于海量的数据处理, 此算法大大提高了关联规则挖掘的效率, 可以快速而且准确的的挖掘出关联规则结果. 但是, 当数据量异常庞大的时候, fp-tree 的构造又相当麻烦, 占据巨大的内存, 增加了时间复杂度, 严重影响了下一步的频繁模式挖掘. 针对数据量大的效率问题, 本文采用剔除一些复杂节点的方法, 首先扫描一遍数据库, 找出偏离点并将其删除, 然后对剩余缩减后的数据进行挖掘.

1 物流营销决策问题

在世界上, 物流活动具有非常遥远的历史, 在我国古代就有跨国界物流, 如著名的丝绸之路. 数百年来, 在频繁的物流活动中人们积累了大量的经验, 却一直没有形成一定的系统理论. 一直到二战以后, 世界上才逐步形成系统的物流理论, 物流活动才开始真正形成.

美国物流学者 Bernard Lagat 曾说, 物流活动是由于地区产品的剩余导致的地区间产品的交换, 导致物流业的出现和发展[4]. 旺盛的生产力推动着物流业的发展, 物流伴随着经济的不断发展而发达, 但是, 物流活动在各个环节都会产生大量的、结构复杂的数据, 很大程度上提高了物流企业决策的复杂度, 使决策不确定性、模糊性和随机性加剧, 同时为物流业的决策层提供了一个很好的决策平台. 针对这一问题, 已有部分先驱者把数据挖掘技术与物流信息系统结合起来服务于物流决策问题, 徐鑫涛针对仓储的合理化安排、制定合理的库存策略、客户价值分析等问题, 提出用数据挖掘技术解决物流仓储决策问题[5].

2 关联规则算法

2.1 关联规则概念和理论

设 $I=\{I_1, I_2, I_3,.....,I_m\}$项的集合. 事务数据库的集合是数据集 D, 其中每个事务 T 是项目的集合, 使 $T\subseteq I$. 每一个事务称作 TID, 事务 T 包含一个项目集 X 当且仅当 $X\subseteq T$, 一个关联规则就是形如 $X\Rightarrow Y$ 的逻辑蕴涵式, $X\subset I$, $Y\subset I$ 且 $X\cap Y=\varnothing$.

(1) 支持度: 设项集 X 和 Y, 其并集 $A\cup B$ 在所有事务 D 中所出现的概率.

$Support(X\Rightarrow Y)=P(X\cup Y)= Support(X\cup Y)=S$.

(2) 置信度: 在事务 D 中既出现了项集 X 又出现了项集 Y 的概率, $Confidence(X\Rightarrow Y)=P(Y|X)= Support(X\cup Y)/Support(X)=Z$. 关联规则的挖掘算法分为两个步骤[6]:

(3) 频繁项集: 所有项集均满足最小支持度阈值.

(4) 产生规则: 从上一步发现的频繁项集的规则里抽取大于置信度阈值的规则, 即强规则.

2.2 频繁项目集

设 $I=\{I_1,I_2,I_3,.....,I_n\}$是一个项目集合, 事务数据库 $T=\{T_1,T_2,T_3,.....T_n\}$是由一系列具有唯一标识的 TID 事务组成, 每个事务 $T_i(i = 1,2,....n)$ 都对应 I 上的一个子集.

定义 1. 设 $I_1\subseteq I$, 项目集 I 在数据集 T 上的支持度是包含 I_1 的事务在 T 中所占的百分比, 即 $support(I_1) =\|\{t\subseteq T|I_1\subseteq t\}\|/\|T\|$.

定义 2. 如果用户给定的最小支持度 minsup 小于等于项目集的支持度则该项目称为频繁项目集[7].

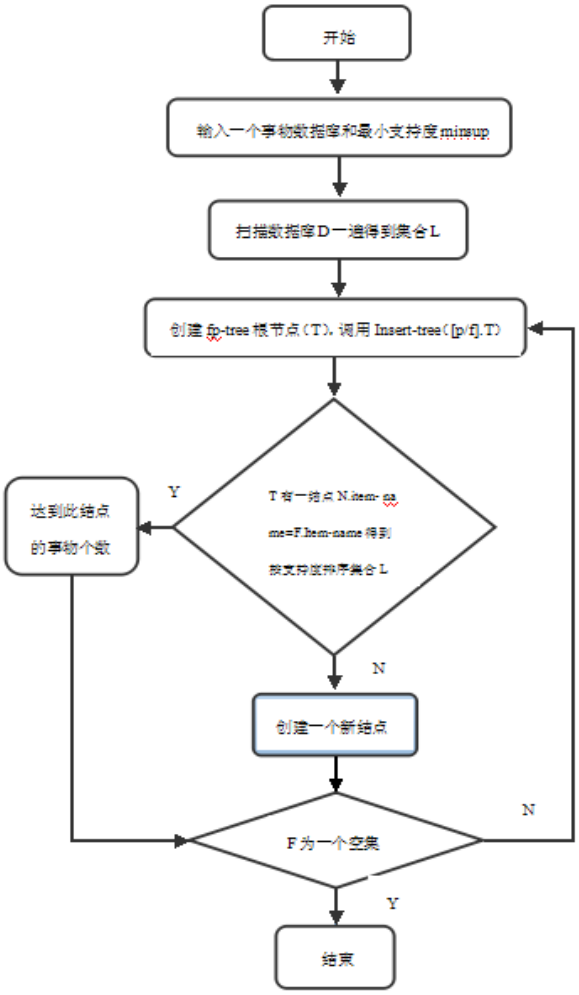

图 1 fp-tree 的构造流程图

2.3 fp-growth 算法思想[8]

FP-growth 算法, 只需扫描两次数据库, 减轻了对输入输出的压力, 不用候选集模式, 减少了庞大候选集引起的问题, 该算法有效提高了效率. 算法的大致思想是扫描一遍数据库获取全部频繁项目, 由高到低排序这些频繁项目, 然后再次扫描数据库, 构建频繁模式树 FP-tree (FP-tree 上只保留事务中的频繁项目), 最后再从 FP-tree 中获取频繁模式.

2.4 fp-tree 的构建

FP-Growth 算法[9]包含以下两个阶段:

1. FP 树构造阶段:

①扫描事务数据库, 求解 1-项频繁项目集 L, 并对 L 按降序排列.

②创建 FP 树的根节点, 并以“null”标记. 对数据库中每个事务, 先将事务中的 item 按 L 中的顺序排序, 然后再将排序后的每个 item 递归插入到树中. Fp-tree 构造流程图如图 1 所示.

2. FP-Growth 的挖掘阶段:

通过调用 FP_Growth(FP_Tree, null)实现. 该函数流程图如图 2 所示.

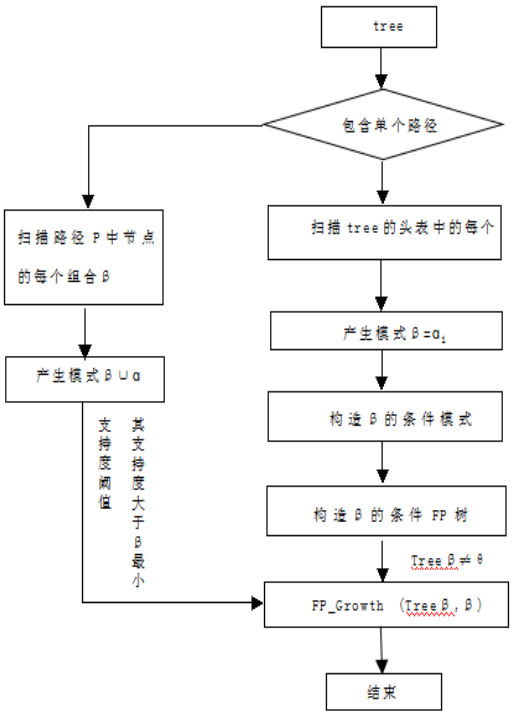

图 2　FP_Growth 函数流程图

FP_Growth (Tree, α) //Tree 为待挖掘的频繁模式树; α为条件模式基, 初值为空

2.5 fp-growth 算法挖掘过程

根据北京物资学院薛华[10]的介绍, fp-tree 的构建算法描述如下:

第一步, 扫描事务数据库, 得到频繁数据项的集合 F 及其支持度. 对其中的元素按支持度进行降序排列, 得到频繁数据项列表.

第二步, 创建以“null”为根结点 r 的 FP-Tree. 针对数据库中的所有事务, 从第三到第五步进行重复执行.

第三步, 按 F 中的次序排列事务数据库中的频繁数据项. 排序后的频繁数据项表记为[p|P], 其中 p 是第一个频繁数据项, 是剩余数据项目列表.

第四步, 调用函数 Insert-Tree([p|P],T), 首先由根结点 T 始, 若 p 有子结点并且 p.item-name 等于 P.item-name 的值, 则结点 N 的计数就增加 1; 否则重新建立一个新的结点设置计数值为 1, 并将其父结点 r 相连, 然后凭借结点的链式结构将值域 Item.name 一样的结点和新结点 V 链接.

第五步, 如果频繁数据项表 P 为非空, 则递归调用函数 Insert-Tree (P,N).

3　算法优化

(1) 扫描一遍事务数据库剔除离异无用点, 生成频繁项集, 对支持度相同的项做特殊处理, 比较该项与前后项的支持度, 较大者先排序, 如: 设前项为 X, 中间项为 Y、Z 且支持度相同, N 是后项; 在降序排列 Y, Z 时, 先计算 support(XYN)和 support(XZN), 支持度最大的排在前面.

(2) 构造局部 fp-tree, 生成局部条件模式树, 挖掘出局部频繁项目集 A 和最大频繁模式树 M.

(3) 整合局部频繁项目集 A, 生成全局频繁项目集 B, 对小于最小支持度频繁项的进行删除.

(4) 挖掘最大频繁模式树 M, 得出关联规则 R.

(5) 将得到的关联规则 R 与项目集 B 进行一一匹配.

(6) 匹配成功则结束, 否则继续循环挖掘直到得到正确的关联规则.

4　改进fp-growth算法对订单信息进行关联规则提取

4.1 数据预处理

数据挖掘过程中, 数据预处理十分重要, 如果处

理不当将造成后续工作无法进行, 严重影响频繁模式树的构建, 而对于试验中海量数据的结构复杂、形式不一, 模糊性强并且占用大量内存, 数据的预处理阶段也是相当麻烦的. 这就需要对数据中的大量偏离点或复杂点进行清除, 在不影响有效数据的情况下, 极大的缩小了实验数据量.

首先, 扫描一遍数据库将数据中存在的大量偏离点或复杂点进行剔除, 这样大大减少了无用数据的干扰, 然后对数据库中剔除无用点后的海量历史数据进行清洗和转换, 达到算法运行所要求的数据格式. 在本系统中, 主要是要挖掘烟草物流中心近 3 个月内被订购的各产品之间的关系. 订单表中的数据提取和转换为将客户订单转换成事务表(如表 1)格式所示.

表 1 事务表

订单名	下单产品名
T20140321	$t_1,t_3,t_5,t_8,t_{20},t_{21},t_{23},..$
T20140724	$t_2,t_6,t_7,t_4,t_{11}...$
T20141234	$t_1,t_5,t_{12},t_{34},t_{25},t_9...$
T20145213	$t_1,t_4,t_8,t_{10},t_7,t_{33},t_{29},t_5...$
T20140311	$t_2,t_5,t_7,t_9,t_{12},t_{14},..$
T20140202	$t_6,t_3,t_5,t_6,t_2...$
...	....

随着烟草产品的不断销售, 数据库里的数据不断在更新, 为了能及时而准确的从这些历史销售数据中提取出对营销决策有用的信息, 需要对各个时间段的销售数据进行处理, 在数据预处理模块, 用户需要输入订单的起始和截止时间, 系统将数据库中这段时间内的订单数据自动转换生成事务数据作为算法的输入项.

数据转换过程如下: 读取某个时间段的客户订单, 以订单名作为事务标记, 然后读取此订单下的所有产品, 转换成图 3 中的格式, 所有订单数据转换完成后, 将得到关联规则挖掘所需要的事务数据库.

4.2 关联提取

事务数据的准备阶段完成之后就要进行关联规则的提取, 对于 FP-Growth 算法而言还需要输入以下几项: 订单事务数据、minsup(最小支持度)、minconf(最小置信度), 算法输出: FP-Tree、全部频繁项集、关联规则. 算法参数的设置需要根据实际的订单数据量, 预期要挖掘的结果以及丰富的经验来确定, 参数设置的太大, 会遗漏有用的规则或者挖掘不出来, 太小又会使大量没有实际意义的规则会被挖掘出来[11], 因此为能达到良好的挖掘效果就需要不断地调整, 对于所挖掘出来的规则, 则需要根据实际的销售情况和以往的销售经验, 来确定有效的挖掘结果.

本实验针对郑州烟草物流中心某 3 个月内的 47.4 万历史订单数据进行了挖掘, 参数设置为 minsup=4, minconf=0.8, 实验结果(如图 3)所示.

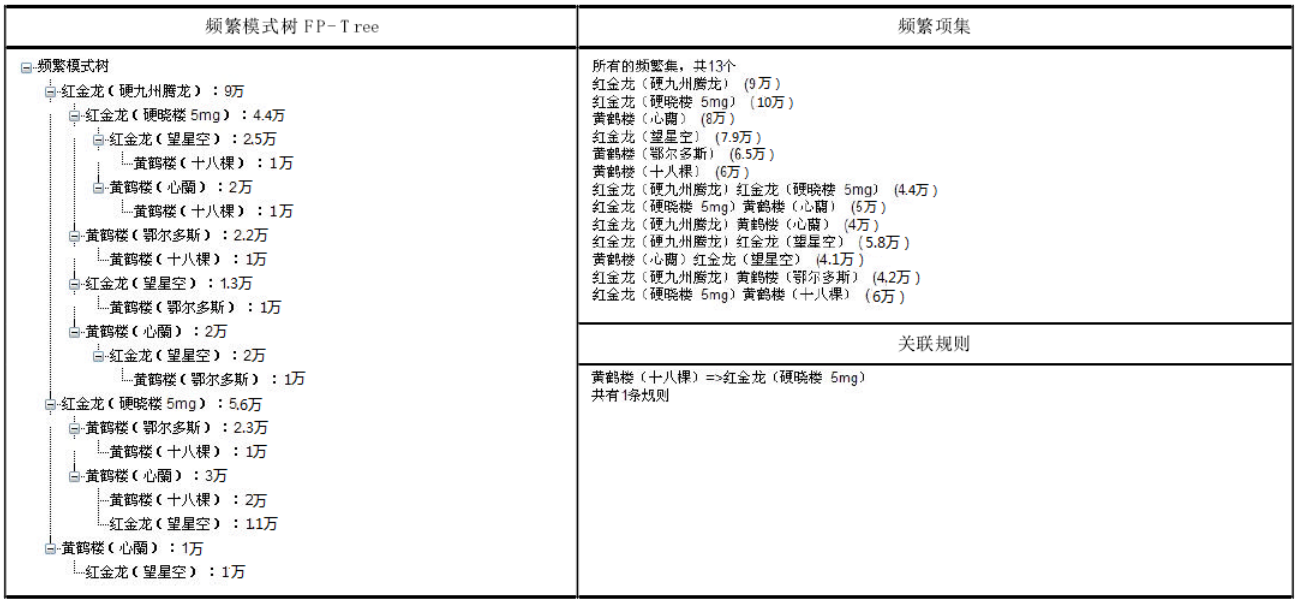

图 3 客户订单关联规则挖掘结果

4.3 关联结果分析

从上面挖掘结果可得出如下规律, 在这一段时期内, 烟草产品“黄鹤楼(十八棵)”和“红金龙(硬晓楼 5mg)”存在一定的规律关系. 当烟草商户在向烟草公司大量订购“黄鹤楼(十八棵)”时还倾向于订购红金龙(硬晓楼 5mg)”这种产品. 当参数变化时, 所挖掘出来的关联规则的数量和规律也因此而不同, 其中有的关联规则是有意义的, 而有的则是无没有意义的, 这就需要进行多次实验来找出一个合适的参数, 同时需要丰富的销售经验和相关知识, 对挖掘结果的有效性进行准确的判断, 当所抽取数据的时间段不同时, 也同样能够挖掘出相应的不同规则.

4.4 制定营销策略

在海量数据中挖掘出其中潜在的高价值信息对烟草销售企业将是至关重要, 烟草物流公司根据这些潜在信息制定相应的营销策略对公司的发展前景尤为重要. 本实验挖掘出的关联规则关系为烟草商户在大量订购“黄鹤楼(十八棵)”后还倾向于订购“红金龙(硬晓楼 5mg)”这种烟草产品, 根据这层关联关系可以利用推荐系统, 在烟草商户网上订购“红金龙(硬晓楼 5mg)”时提醒推荐订购“红金龙(硬晓楼 5mg)”, 不但节省了客户订购时间, 而且更人性化的服务于客户, 提高了客户的满意度, 在最大程度获取利润的同时也维持了客户的数量. 烟草物流公司在配送时, 注意两产品的摆放顺序, 这样大大节省了配送时间. 由于数据量比较庞大而且处于不断更新状态中, 在实时更新数据的同时需要不断采集数据进行关联规则挖掘, 以便准确的挖掘出潜在的有用的信息.

5 结语

对于烟草物流公司而言, 数据的不断更新将产生大量信息数据, 实验表明改进后的 fp-growth 算法更适合处理比较庞大的数据量, 使内存的占用量小, 且响应速度更快, 改进后的关联规则算法可以有效地提升烟草物流企业在市场的反应速度, 满足各种销售商的需求, 能够帮助烟草物流公司获得较强的竞争优势. 降低物流企业的运营成本提高决策效率, 在未来必将有着广阔的市场前景.

本系统经过长期的试用, 能较好的找出不同销售时期对应的不同产品之间的销售规律, 为烟草物流公司制定相应的营销策略具有很好的参考价值. 以该算法为基础设计的关联规则挖掘系统, 数据量较小的情况下可以达到秒级响应, 数据量很大的情况下也能比较快速而准确的得出挖掘结果, 其运行时间范围可以接受, 具有较好好的实用价值.

参考文献

1 刘世平.SQL 数据挖掘技术在物流企业管理中的应用.物流技术,2013,32(3):209–211.

2 肖娟.数据挖掘在物流业的应用综述.统计与决策,2013,383(11):95–97.

3 魏峰.基于聚类的关联规则挖掘算法研究[硕士学位论文].杭州:浙江工业大学,2012.

4 王戈秋.物流管理与市场营销.物流商论,2012,(1):181–182.

5 杨艳霞,杨丽华,张伟丰.基于 FP-Growth 算法的卷烟产品销售规律挖掘研究.科技创业,2013,(4):31–33.

6 晏杰,亓文娟.基于 Apriori & FP-growth 算法的研究.计算机系统应用,2013,22(5):122–125.

7 章志刚,吉根林,等.一种基于 FP–Growth 的频繁项目集并行挖掘算法.计算机工程与应用,2014,50(2):103–106.

8 周钦良,李玉忱,公爱国.一种新的高效生成 FP-Tree 条件模式基的算法.计算机应用,26(6).

9 Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM SIGMOD Record, 2000, 29(2): 1–12.

10 薛华.数据挖掘在物流客户关系管理中的应用研究[硕士学位论文].北京:北京物资学院,2012.

11 杨艳霞,张伟丰.卷烟产品销售规律挖掘算法,2013,(1):219–220.