

关联规则数据挖掘算法及其在智能物流中的应用分析

戴小廷

(福建工程学院 交通运输系, 福州 350108)

摘要:关联规则数据挖掘算法是近年国内外学者关注的一个课题。国内外学者为了提高挖掘的效率对其做了不断的改进研究。在关联规则数据挖掘算法的原理分析的基础上,对国内外该算法的研究现状做了认真分析,并分析了其在商业、医疗、保险、通讯、工业、农业等领域的典型应用,提出将不同的算法应用于实际挖掘工作中是未来进一步深化研究的方向。结合当前物流企业数据分析现状与需求,分析了关联规则数据挖掘在智能物流中的具体应用,为物流企业改善物流营销服务、优化物流路径等提供决策支持。

关键词:关联规则;数据挖掘;智能物流;决策

中图分类号: TM 73 **文献标志码:** A **文章编号:** 1671-1807(2014)02-0113-04

关联规则数据挖掘是数据挖掘中是一个重要的课题,近年来一直是信息技术与管理领域研究的热点。从大量的数据集中找出数据项之间的关联关系,即某一项或某几项数据的出现在一定的置信度下另一项数据或另几项数据出现的可能性,也即数据之间的这种关联关系就是关联规则,这种关联包括简单关联、时序关联、因果关联,一般来说,寻找隐藏在数据中的这类关联规则特别是高置信度和可信度的规则能为特定用户提供有价值的信息,从而支持其决策过程。在关联规则挖掘的方法中,最经典的算法就是 Apriori 算法,除此之外还有 Patition 算法、完全频繁项集挖掘、频繁闭项集挖掘、最大频繁项集挖掘算法以及一些新的频繁项集挖掘算法,因需要大量候选集开展挖掘致使大多数寻找频繁项集算法效率不高。

最近几年就改进挖掘效率,许多学者开展了相关研究,取得了一些成果,但总体而言,这些改进的算法还是都有各自的优缺点,应用中有不同的限制条件,比如在处理大量的数据方面算法效率还不够高、既有算法对于流数据的挖掘还需改进、寻找并行算法中的有效平衡负载点等问题都有待于进一步的研究等。

1 关联规则数据挖掘算法原理

数据挖掘技术自 20 世纪 80 年代出现以来,经过众多学者的努力,理论和方法已日趋成熟,至今它已形成了一系列的挖掘算法,特别是在人工智能、机器

学习、数据仓库等技术的支持下,不断产生各种优化和创新的算法,为数据挖掘工作的开展奠定了较坚实的基础。

在众多的数据挖掘算法中,关联规则数据挖掘算法是一种产生较早且比较基本的方法,该算法简单、适用,可操作性强,其基本原理就是设定数据出现频率的支持度以及关联规则的可信度,采取特定的算法从已知的事物数据集中找出符合支持度要求的频繁项集,然后再频繁项集里面应用剪枝等策略获取符合可信度要求的关联规则。在整个关联规则的挖掘过程中,比较关键的是频繁项集的寻找,如何快速准确的找出所有符合设定条件的频繁项集是关联规则数据挖掘算法研究的重点。

2 关联规则数据挖掘算法研究进展

2.1 关联规则数据挖掘算法国内外研究综述

近二十年来国外许多数据挖掘领域的研究人员投入了大量的时间和精力,深入研究了关联规则的算法,其中 Agrawal 等人于 1993 年提出的 Apriori 算法就是其中最具代表的成果,随后众多学者又在此基础上提出了一些改进,目的在于提高算法的效率,从而改进数据挖掘的效率。这方面的研究是目前数据挖掘领域研究的热点之一。

Apriori 算法是挖掘布尔关联规则频繁项集的一种非常经典的算法^[1]。Han Jia.pei 等人在 Apriori

收稿日期:2013-12-17

基金项目:福建工程学院 2011 年度校科研发展预研基金(GY-Z11077)

作者简介:戴小廷(1972—),男,江西宜春人,福建工程学院,副教授,博士,研究方向:信息管理与信息系统,供应链管理
等。

算法的基础上提出了直接生成频繁项集的算法 FP-growth,在数据集相当密集或支持度阈值较低时其挖掘性能更加出色,弥补了原算法在处理长模式和密集数据集时性能低下的不足^[2]。FP 树法并不产生候选项集,关联规则的算法可按照需不需要产生候选项集的做法分为两类,以 FP(频繁模式)树法与 Apriori 为代表。Park 等人提出将哈希表结构用于关联规则的 DHP 算法能有效进行数据集删减和过滤由 k-项集连接产生的候选集,有效地减少了候选集的数目,但由于创建 hash 表和每次迭代时将更改后的数据集写入磁盘增大了数据库空间耗用^[3]。为了改进循环规则的表达能力,贝尔实验室的 Ramaswamy 等人提出挖掘符合日历的关联规则(calendric association rules)算法有效进行了市场货篮分析。另外为了加快关联规则的挖掘速度和提高挖掘效率,综合考虑计算、通信、内存的利用和并发之间的协调关系一些学者设计了关联规则挖掘的并行算法,比如 Rakesh Agrawal 等人提出的 Count 分布、Data 分布和 Candidate 分布算法,Jong Soo Park 等人提出的 PDM 算法,Eui-Hong Han 等人提出的智能数据分布算法 IDD 和混合分布算法 HD,David Wai-Lok Chenung 等人提出的 DMA、FDM 和 APM 算法等。

近年来,国内学者在学习引进国外的数据挖掘技术的过程中,也开展了深入的研究,对关联规则数据挖掘算法做了各种改进研究,推动了该挖掘方法的发展。如刘德喜等人提出了 FIS-ES 算法。胡学钢等提出基于 PCL 模型的频繁项集求解算法,改善了频繁项集挖掘算法的时空性能^[4]。战立强对 FP-tree 进行了改进,提出了新的搜索与剪枝策略,使该算法的搜索效率得到提升^[5]。陈江平等运用概率来对 Apriori 算法的频繁项集的取得进行优化,减少了候选集,从而提高得到频繁项集的效率^[6]。文蓉等针对 Apriori 算法需要重复地扫描数据库的缺陷,提出基于事务地址索引表来约简事务的 Apriori 优化算法,通过建立事务地址索引加快搜索定位,提高算法执行效率^[7]。刘星沙等在生成候选项集时将低于计算支持度的候选项集中的事务删除,提出了加快产生频繁项集的速度的算法^[8]。谢宗毅所提出的 S_Apriori 算法也提高了算法的效率^[9]。

总之,通过国内外相关研究文献发现,经过众多学者的努力,关联规则数据挖掘算法自产生以来,经过不断的改进和优化,在不同的侧重点上对原有算法的效率有所提升。

2.2 关联规则数据挖掘算法的典型应用

在过去的几年里,伴随着关联规则数据挖掘算法的丰富和发展,该算法在实际领域有了较多的应用,在商业、医疗、保险、通讯、工业、农业等领域都有开展具体应用,取得了一定的成效。如陆璐将关联规则的数据挖掘应用在集装箱追踪系统中,找出客户发货去向之间,以及发货种类与发货去向之间的关联关系,为其开展物流市场营销,提供科学、低风险的决策依据^[11]。刘尚辉等将 Apriori 关联规则在甲状腺结节病案分析中进行了应用,通过挖掘应用形成 251 个强关联规则,这些强关联规则中蕴含着甲状腺结节发病与性别、年龄、甲状腺功能及血常规指标等因素之间的关联关系。通过该方法有利于将目前医院中大量医疗病例通过数据挖掘方法转变为实用的新知识,为医院大量病案信息提供了新的研究方法,具有一定的应用价值^[11]。邱凤翔等将关联规则数据挖掘算法应用在火电厂不同机组测点的关联关系的分析上,论证了这方面应用的可行性^[12]。张广群等应用空间关联规则在林业信息系统中进行了研究,提出了从小班、行政区域、河流和公路中获取空间关联规则的方法,并以某县级市为例,获得了初步的效果^[13]。熊正德等运用 KMV 模型计算上市公司违约距离作为衡量上市公司信用风险的指标,并且使用 Apriori 算法对上市公司信用风险进行关联规则挖掘,真实反映上市公司之间信用风险传染效应,相关公司或投资者可以根据关联规则及时调整策略规避风险^[14]。

2.3 关联规则数据挖掘算法的进一步发展分析

国内外众多学者从不同的角度提出了各种优化算法,尽管这些算法各具优点且挖掘的性能和效率均明显高于传统的算法,但总的来说,各自还是有些缺点,算法仍较复杂。如何针对不同的算法在实际挖掘工作中加以应用,从而提高挖掘工作效率,给实际工作带来好处,还是数据挖掘工作者面临的难题之一,还需要更多的深入研究。

3 关联规则数据挖掘算法在物流企业智能物流中的应用前景分析

3.1 物流企业对数据分析的新要求

在物流企业中,随着计算机信息技术的发展和多年的信息化建设,企业已建立了运行稳定的各类信息系统,在经营管理中发挥着重要的作用。但随着信息系统长期运行,物流企业又面临着一些新问题,如合理存储 MIS 的历史数据及如何有效对海量的日积月累下来的历史数据进行分析,为运输管理、仓储管理、供需分析、人力资源管理各方面提供更好为经营决策服务等。特别是在当前物联网建设大环境下,物流企

业要想真正实现智慧物流,都离不开数据挖掘技术的支持,研究数据挖掘技术在物流管理过程中的具体应用,是当前物流企业知识管理工作中的重点和难点之一。而纵观国内外的已有研究,关联规则数据挖掘的很多研究还停留在理论研究和实验验证上,在物流企业中的应用还处于初步阶段,开展相关的应用研究较少,且不够系统具体,所以本文将结合物流企业实际,探究算法与物流企业具体数据挖掘工作实际的契合点,结合实际应用来验证和改进算法,寻求理论与实际的良好结合。

3.2 应用关联规则数据挖掘开展物流企业智能物流数据分析

从前面应用分析得知,关联规则挖掘的应用非常广,而国内将关联规则数据挖掘应用于物流企业的应用研究开展较少,在当前物联网建设大环境下,一方面物流企业面临需要不断提高物流效率、降低物流成本、提高客户满意度、满足越来越多的个性化需求的外在要求。另一方面,在物流企业,随着网络技术、RFID、条形码与各类信息管理系统的应用推动,在物流信息的采集自动化、数据处理便捷快速等方面取得了长足的进步,经过多年的建设,积累了海量的数据,这些数据中隐含了大量潜在的、有用的知识,通过应用数据挖掘的技术,将挖掘到的规则与物流管理各方面有机地结合,将能为企业经营决策提供科学有力的支持。通过各类数据挖掘技术的应用研究开展将丰富和拓展物流企业数据分析手段,使物流企业能深入、全面地分析业务数据,从各种不同的角度审视业务的经营情况,改善物流企业数据分析的准确性和及时性,科学地、快速地指出存在的问题、隐患并能发现对于企业至关重要的变化趋势,比如找出库存物品和成本之间的关联关系、采购成本差异与原因的关联关系、开展商品退货的原因分析、寻找运输路径与成本之间的关系等等,通过挖掘,发现企业实际存在的问题,为科学决策做出坚强的数据支持。同时为企业智能物流系统的设计提供有益的帮助。

目前物流企业的数据挖掘技术应用还多停留在对某一领域点的研究上,系统性不够,未能综合考虑全盘业务,且对数据挖掘系统应用设计以及原有信息系统的集成方面研究还不够,所以实际应用中对算法还需要考虑多方面的影响并不断地加以改进,同时如何让数据挖掘效率更高、结果展示更加简单、科学、易于理解等问题都还有待进一步深入研究。

3.3 基于关联规则的物流决策应用实例

选取福州某第三方物流企业部分货物流向事务

数据库数据元组作为训练集,归类处理后的部分数据见表1所示。采用Apriori算法对物流路径数据开展分析,挖掘出频繁出现的路径信息,并由频繁项集寻找出关联规则,即货物流向关联性及其可能出现的变化。

表1 福州某物流企业部分业务数据

业务日期	运单编号	物流路径	运输方式
2013-05-04	D010711	温州,上海,南京	集装箱
2013-05-04	D010813	上海,南京,合肥	行包
2013-05-05	D010712	温州,上海	集装箱
2013-05-08	D010715	上海,南京	集装箱
2013-05-08	D010716	温州,上海,南京,合肥	集装箱
2013-05-09	D010722	温州,常州	集装箱
2013-05-12	D010723	温州,南京	集装箱
2013-05-12	D010814	温州,上海,南京,常州	行包
2013-05-13	D010750	南京,常州	集装箱
2013-05-14	D010762	温州,南京	集装箱

本文采用Microsoft SQL Server 2005中的Analysis Services工具开展关联数据挖掘,设定最小支持度为20%,最小置信度阈值为80%。其主要工作原理是首先通过扫描训练集,对每个候选记数。得到候选1—项集,然后比较候选支持度计数与最小支持度计数得到频繁1—项集,再由频繁1—项集产生候选2—项集,比较候选支持度计数与最小支持度计数(20%),得频繁2—项集,同理得到频繁3—项集,直到找到所有频繁项集。经过挖掘,在最小置信度阈值80%条件下得到如下的关联规则:

- 南京∧合肥→上海 置信度 100%
- 上海∧合肥→南京 置信度 100%
- 合肥→上海∧南京 置信度 100%

上述关联规则反映了客户发货去向之间的关联关系,能为物流企业改善物流营销服务、优化物流路径等提供决策支持,最终提高物流绩效。

4 总结

关联规则数据挖掘方法经过国内外众多学者的努力,产生了众多的改进算法,在算法的效率上取得了不少的改善,各自有自身的优缺点,在企业实际工作中的应用也已取得了一定的成效,但还需要不断完善。特别是在物流企业,基于关联规则数据挖掘的企业智能物流数据分析工作还处于起步阶段,结合物流数据特点开展具体的数据挖掘工作还需要进一步的深入研究。

参考文献

[1] AGRAWAL R, SRIKANT R. Fast Algorithm for Mining As-

sociation Rules in Large Database[Z]. Proc. 1994 Int Conf. on VLDB, Santiago, Chile, 1994:487—499.

[2] HAN J, PEI J, YIN Y. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53—87.

[3] PARK J H, KANITKAR V, DELIS A. Logically Clustered Architectures for Networked Databases[J]. Distributed and Parallel Databases, 2001, 10(2): 161—198.

[4] 刘德喜, 何炎祥, 邢显黎. 一种新的频繁项集挖掘算法[J]. 计算机应用研究, 2007, 24(2): 17—19, 62.

[5] 战立强, 刘大昕. 频繁项集快速挖掘算法研究[J]. 哈尔滨工程大学学报, 2008(3): 266—271.

[6] 陈江平, 傅仲良, 徐志红. 一种 Apriori 的改进算法[J]. 武汉大学学报, 2003(2): 94—98.

[7] 文蓉, 李仁发. 一种优化的 Apriori 算法[J]. 计算机系统应用, 2008(1): 94—96.

[8] 刘星沙, 谭利球, 熊拥军. 关联规则挖掘算法及其应用研究[J]. 计算机工程与科学, 2007(1): 83—85.

[9] 谢宗毅. 关联规则挖掘 Apriori 算法的研究与改进[J]. 杭州电子科技大学学报, 2006(6): 78—82.

[10] 熊正德, 冷梅. KMV 和 Apriori 算法在上市公司信用风险传染中的应用[J]. 湖南大学学报, 2010, 24(3): 58—61.

[11] 陆璐. 基于关联规则的数据挖掘在集装箱追踪系统中的应用[J]. 铁路计算机应用, 2008, 17(8): 25—28.

[12] 邱凤翔, 徐治皋, 司凤琪, 等. 基于主元分析的关联规则挖掘及在电厂中的应用[J]. 电力系统自动化, 2009, 33(16): 81—84.

[13] 张广群, 方陆明, 汪杭军, 等. 空间关联规则在林业信息系统中的应用[J]. 林业资源管理, 2009(6): 120—122.

[14] 刘尚辉, 王露, 郑德禄. Apriori 关联规则在甲状腺结节病案分析中的应用[J]. 中国卫生统计, 2011, 28(2): 178—179.

[15] 王爽, 王国仁. 基于滑动窗口的 Top-K 概率频繁项查询算法研究[J]. 计算机研究与发展, 2012, 49(10): 2189—2197.

Analysis of Association Rule Mining Algorithm and Its Application in Intelligent Logistics

DAI Xiao-ting

(Department of Transportation, Fujian University of Technology, Fuzhou 350108, China)

Abstract: The association rules mining is one of problems attracting scholars of the world in recent years. In order to improve the mining efficiency, scholars of the world continuously study. This paper introduces the principle of association rules mining algorithm and analyzes the domestic and foreign research status of the algorithm, and analyzes its typical application in commercial, medical, insurance, communications, industry, agriculture and other fields, the algorithm is applied in different actual mining work is a further study of the future direction; combined with the analysis of current situation and demands of logistics enterprise data, analysis of association rules of data mining application in intelligent logistics, provide decision support for optimizing logistics path to improve logistics marketing service of logistics enterprises.

Key words: association rules; data mining; intelligent logistics; decision making

(上接第 44 页)

[8] 鲁明勇, 彭延炼. 我国旅游研究应用博弈论的现状和问题[J]. 科技和产业, 2006 (8) : 66—69.

[9] 郭华. 乡村旅游社区利益相关者研究: 基于制度变迁的视角[M]. 广州: 暨南大学出版社, 2010: 79—82.

The Game Analysis of the Tourism Industry Stakeholders

ZHANG Yan, LONG Zu-kun

(Tourism Management Institute of Xiangtan University, Xiangtan Hunan 411105, China)

Abstract: The further development of the tourism industry, which leads to the game between stakeholders tourism enterprises also gradually revealed, in order to better maximize their own interests. The three types of government, the scenic area residents and tourists analyzes the game between stakeholders and tourism enterprises, seek for the countermeasures of sustainable development, of the many stakeholders' interests so as to promote the development of the tourism industry to achieve a state of interests balance, symmetry in liability.

Key words: tourism industry; stakeholders; game theory