

基于关联规则的搭配进货系统的研究与实现

王 妍,王丽君,方 芸

(曲阜师范大学 计算机科学学院,山东 曲阜 276826)

摘 要:为了解决商品进货无关联的现状,找到商品间的关联规则,更好地进行商品的搭配进货,从而提高进货效率,文中引入了关联规则的思想,并利用规则进行了商品关联规则的挖掘。在分析了关联规则挖掘的算法后,将其应用到超市商品数据库中,利用关联规则挖掘出大量数据中项集即商品之间的相互关联,并抽取有价值的商品关联规则,利用支持度和平衡度这两个度量概念,优化出强规则集,并用这一思想成功设计了 PLM 即产品全生命周期管理中的搭配进货系统。

关键词:关联规则;数据挖掘;搭配进货

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2012)01-0137-03

Research and Realization of Restocking System Based on Association Rules

WANG Yan, WANG Li-jun, FANG Yun

(Computer Science Institute, Qufu Normal University, Qufu 276826, China)

Abstract: In order to address the status of commodity stock that are not associated, find associations rules of commodities and matching of goods stock, so as to improve the efficiency of stock, introduced the idea of association rules and used rules for commodity association rule mining. In this paper, algorithm of association rules mining is analyzed and applied to the database of supermarket goods. The valuable association rules of supermarket goods were extracted, then the strong rules were extracted by using the two metrics concept of support and balance. Finally, all strong rules were applied to the restocked system of PLM.

Key words: association rules; data mining; restocking system

0 引 言

关联规则可以挖掘出大量数据中项集之间的相关联系,在数据挖掘中是一个重要的课题,其中,关联规则挖掘的一个典型例子是超市购物篮分析,即通过关联规则的研究,发现交易数据库中不同商品(项)之间的联系,找出顾客购买中的经常性行为模式,如购买了某一商品的顾客,经常同时也购买其他商品^[1]。分析结果可以应用于超市商品货架的布局、货存的安排以及根据购买模式进行搭配进货等^[2]。

1 关联规则概述

1.1 关联规则定义

为了更好地理解关联规则,首先介绍与关联规则算法相关的概念,以及关联规则挖掘算法。

形式上下文:一个形式上下文是一个三元组 $(G,$

$M, I)$ 。其中, $G = \{g_1, g_2, \dots, g_n\}$ 是有限的对象集合, $M = \{m_1, m_2, \dots, m_k\}$ 是有限的属性集合, $I \subseteq G \times M$ 上的二元关系。如果 $(g, m) \in I$, 则称对象 g 有属性 m ^[3]。

操作 ∇ 的定义:给定集合 $A \subseteq G, A^\nabla = \{m \in M \mid \forall x \in A (x, m) \in I\}$ 。表示集合 A 中的所有顾客都购买的商品集合。同样,对于 $B \subseteq M$, 有 $B^\nabla = \{g \in G \mid \forall y \in B (g, y) \in I\}$, 表示购买了集合 B 中所有商品的顾客的集合(注释: $A^{\nabla\nabla}$ 表示 $(A^\nabla)^\nabla$, 即先对 A 进行一次 ∇ 操作, 然后对 $A^\nabla \subseteq M$ 再进行一次 ∇ 操作。同理, $B^{\nabla\nabla}$ 表示 $(B^\nabla)^\nabla$)。

形式概念:对于序对 (A, B) , 若满足以下两个条件: 1) $A \subseteq G, B \subseteq M$; 2) $A^\nabla = B, B^\nabla = A$, 那么称 (A, B) 是 (G, M, I) 上的一个形式概念, 简称为概念^[4]。 A 是形式概念的外延, B 是形式概念的内涵。概念层次越高, 说明概念的外延越大, 内涵越小; 概念层次越低, 概念的外延越小, 内涵越大。

伪内涵:称 $B \subseteq M$ 是 (G, M, I) 的一个伪内涵, 如果满足条件: 1) $B \neq B^{\nabla\nabla}$; 2) $\forall P \subseteq B (P^{\nabla\nabla} \subseteq B)$, 其中 P 是伪内涵。

收稿日期:2011-06-22;修回日期:2011-09-27

基金项目:山东省软科学项目(2010RKGA1053)

作者简介:王 妍(1980-),女,山东曲阜人,实验师,硕士研究生,研究方向为人工智能、软件工程、数据挖掘。

规则:如果 $B \subseteq M$ 是一个伪内涵,则称 $B \rightarrow B^{\vee\vee} \setminus B$ 是一条规则^[5]。

集合间的字母序关系:设 $A, C \subseteq G$, 称 A 是按字母序小于集合 C , 如果能区分集合 A 和 C 的最小元素在集合 C 中。形式上, $A <_l C: = \exists i \in \mathbb{N} (A \cap \{1, 2, \dots, i-1\} = \emptyset \wedge C \cap \{1, 2, \dots, i-1\} \neq \emptyset)$ 。

比如, G 是离散的有限集合, 可将 G 写成下面的形式: $G = \{1, 2, \dots, N\}$ 。若 $A = \{3\} \subseteq G, B = \{3, 4\} \subseteq G$, 那么 $A <_l B$, 因为存在 $4 \in B \setminus A$, 使 $(A \cap \{1, 2, 3\} = B \cap \{1, 2, 3\}) = \{3\}$ 。

集合闭操作: 给定集合 B , 定义闭操作 Γ^* 如下: $B^{\Gamma^*}: = B \cup (\cup \{C \mid A \rightarrow C \in \Gamma, A \subseteq B\})$; $B^{\Gamma^* \Gamma^*}: = B^{\Gamma^*} \cup (\cup \{C \mid A \rightarrow C \in \Gamma, A \subseteq B^{\Gamma^*}\})$, 继续下去能发现一个集合 $\Gamma^*(B)$ 满足 $\Gamma^*(B) = \Gamma^*(B)^{\Gamma^*}$, 其中, Γ 是规则的集合。

最小伪内涵: 如果 $B \subseteq M$ 是一个伪内涵, 则 $\Gamma^*((B \cap \{1, 2, \dots, i-1\}) \cup \{i\})$ 是按字母序关系 B 后面的最小的伪内涵或概念, 其中 i 满足 $B <_l \Gamma^*((B \cap \{1, 2, \dots, i-1\}) \cup \{i\})$ 。

判断谓词 BT: 当规则 $B \rightarrow B^{\vee\vee} \setminus B$ 为真时其值为真, 否则为假, 可通过人机交互的方式询问是否为真^[6]。

1.2 关联规则算法

Agrawal 等于 1993 年首先提出了挖掘顾客交易数据库中项集间的关联规则问题, 以后诸多的研究人员对关联规则的挖掘问题进行了大量的研究。这里根据前面提到的定义和定理, 提出基于形式概念分析的关联规则抽取算法^[7]。具体描述如下:

输入形式上下文 (G, M, I) , 其中 $G = \{g_1, \dots, g_K\}$ 和 $M = \{m_1, \dots, m_N\}: = \{1, 2, \dots, N\}$

过程:

- (1) $\Gamma := \{\}$
- (2) $B = \emptyset$
- (3) FOR $i = N$ to 1 DO
- (4) IF $B <_l \Gamma^*((B \cap \{1, 2, \dots, i-1\}) \cup \{i\})$
- (5) THEN $B := \Gamma^*((B \cap \{1, 2, \dots, i-1\}) \cup \{i\})$
- (6) IF BT($B \rightarrow B^{\vee\vee} \setminus B$) = TRUE
- (7) THEN $\Gamma := \Gamma \cup \{B \rightarrow B^{\vee\vee} \setminus B\}$
- (8) ELSE $i := i - 1$
- (9) 对 B 和 Γ 重复步骤 (2) ~ (7)

(10) 如果再没有新的伪内涵出现, 算法结束, 输出 Γ (Γ 就是关联规则集合), 输出形如 $X \rightarrow Y$ 的关联规则集, 其中 $X, Y \subseteq M$ 。

2 基于算法的规则提取

根据上面的概念及算法, 可以从超市商品交易事件中抽取出有价值的规则, 例如购买商品 A 的人同时购买了商品 B 和 C , 这样便可以利用这些规则进行商

品的搭配进货, 从而减少进货过程中的盲目性, 提高进货率^[8]。具体分析过程, 用图 1 表示:

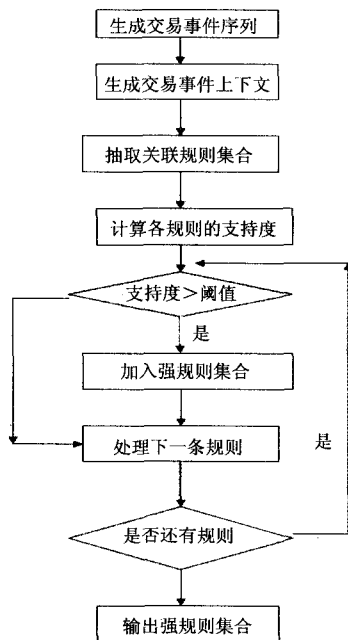


图 1 关联规则抽取过程图

2.1 超市进货系统分析

首先通过超市交易小票, 抽取超市的购买事件, 构成交易事件序列^[9], 如表 1 所示, 其中交易事件一栏表示交易, I_i 表示第 i 次交易事件, 对应的物品集表示第 i 次交易内容, 用 $A \sim I$ 这九个字表示交易商品。

表 1 商品交易事件序列

交易事件	物品集
I_1	$\{A, B, C, D, E, H, I\}$
I_2	$\{A, B, E, F, G, H, I\}$
I_3	$\{A, B, C, D, E, H\}$
I_4	$\{A, E, G, H\}$
I_5	$\{B, E, F, G, H, I\}$
I_6	$\{B, C, D, E, H, I\}$
I_7	$\{A, B, D, I\}$
I_8	$\{A, B, E, F, G, I\}$
I_9	$\{F, G, H, I\}$
I_{10}	$\{C, D, H, I\}$
I_{11}	$\{C, D, E, H\}$
I_{12}	$\{A, D\}$
I_{13}	$\{B, F, G, I\}$
I_{14}	$\{A, G\}$

根据形式上下文的定义, 从表 1 中抽取交易事件上下文, 用表 2 来表示超市交易数据样本, 其中行表示交易事件, 列表示商品, 行列的交叉处表示顾客是否购

表4 规则支持度与平衡度

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12
Support(x)	5/14	5/14	5/14	5/14	2/7	2/7	2/7	5/14	5/14	5/14	5/14	5/15
Bal(x)	1/2	2	2	1/2	2	2	1	2	5	2	1	5

买了此件商品(1表示购买了,0表示没有购买)。

表2 超市交易事件上下文

	A	B	C	D	E	F	G	H	I
I ₁	1	1	1	1	1	0	0	1	1
I ₂	1	1	0	0	1	1	1	1	1
I ₃	1	1	1	1	1	0	0	1	0
I ₄	1	0	0	0	1	0	1	1	0
I ₅	0	1	0	0	1	1	1	1	1
I ₆	0	1	1	1	1	0	0	1	1
I ₇	1	1	0	1	0	0	0	0	1
I ₈	1	1	0	0	1	1	1	0	1
I ₉	0	0	0	0	0	1	1	1	1
I ₁₀	0	0	1	1	0	0	0	1	1
I ₁₁	0	0	1	1	1	0	0	1	0
I ₁₂	1	0	0	1	0	0	0	0	0
I ₁₃	0	1	0	0	0	1	1	0	1
I ₁₄	1	0	0	0	0	0	1	0	0

2.2 规则提取

利用关联规则的获取算法,进行规则提取^[10]。首先,算法开始于空集 \emptyset ,最小的伪内涵是 $\{F\}$, $\{F\}^\nabla = \{I_2, I_5, I_8, I_9, I_{13}\}$, $\{I_2, I_5, I_8, I_9, I_{13}\}^\nabla = \{F, G, I\}$,即 $\{F\}^{\nabla\nabla} = \{F, G, I\}$,所以, $\{F\} \rightarrow \{G, I\}$ 是一条关联规则,即所有购买商品F的顾客同时也买了商品G和I。继续执行算法,得到关联规则集合,如表3所示:

表3 规则集合

规则	内容
Rule 1	$\{F\} \rightarrow \{G, I\}$
Rule 2	$\{E, I\} \rightarrow \{B\}$
Rule 3	$\{B, H\} \rightarrow \{E\}$
Rule 4	$\{C\} \rightarrow \{D, H\}$
Rule 5	$\{A, I\} \rightarrow \{B\}$
Rule 6	$\{A, H\} \rightarrow \{E\}$
Rule 7	$\{D, E\} \rightarrow \{C, H\}$
Rule 8	$\{D, H\} \rightarrow \{C\}$
Rule 9	$\{A, C, D, E, H\} \rightarrow \{B\}$
Rule 10	$\{G, I\} \rightarrow \{F\}$
Rule 11	$\{B, G\} \rightarrow \{F, I\}$
Rule 12	$\{A, B, F, G, I\} \rightarrow \{E\}$

2.3 规则优化

这样产生的规则是利用算法抽取的所有规则,不

是所有的规则都是有价值的,例如规则1和规则10,是两个互补规则,规则1是一个前项决定两个后项,规则10是两个前项决定一个后项,所以规则1要比规则10有价值。所以要对所得的规则进行筛选,提取出真正有用的规则^[11]。

这里利用支持度和平衡度两个概念筛选出所有有价值的规则,生成强规则集。

支持度:定义规则 $X \rightarrow Y$ 的支持度为:

$$\text{Support}(X \rightarrow Y) = \frac{|\{I_i | X \cup Y \subseteq I_i, I_i \in I\}|}{L}$$

例如规则1:包含规则 $\{F\} \rightarrow \{G, I\}$ 的事件有 $\{2, 5, 8, 9, 13\}$,事件总数为14,其支持度就是5/14。

再例如规则5:包含规则 $\{A, I\} \rightarrow \{B\}$ 的事件有 $\{1, 2, 7, 8\}$,其支持度就是2/7。

平衡度:定义规则 $X \rightarrow Y$ 的平衡度为:

$$\text{Bal}(X \rightarrow Y) = \frac{|X|}{|Y|}$$

例如规则1: $\{F\} \rightarrow \{G, I\}$ 的平衡度是1/2;规则5: $\{A, I\} \rightarrow \{B\}$ 的平衡度是2/1。

根据支持度和平衡度的定义,各个规则的支持度和平衡度如表4所示。

这样,支持度越大,平衡度越小,规则的可采纳性就越高,得到的规则越有价值。所有规则中支持度大于规定值,并且平衡度小于规定值的规则可以加入强规则集。规定支持度阈值为2/7,平衡度阈值为1,得到的强规则集为 $\{I1, I4, I11\}$

将这一理论应用在PLM(产品全生命周期)的进货搭配系统中,大大提高了进货效率^[12]。其界面如图2所示:

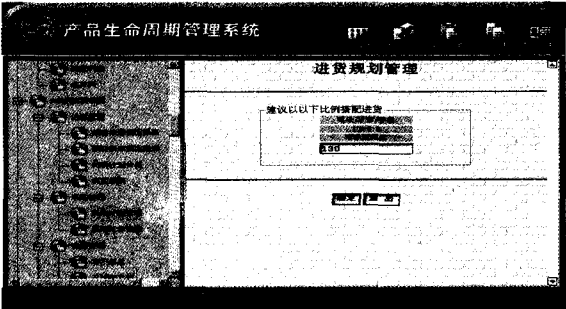


图2 PLM中进货搭配系统界面

3 结束语

文中只是在搭配进货方面做了一些探索,用数据(下转第156页)

测试过程中,每个项目的操作方法、步骤都不同,提示、帮助信息的形式也不一样。软件设计时将它们与项目数据库分离,通过索引加载显示,实现了测试时自动实时地显示提示、帮助信息。即使是第一次使用检查仪的用户也可以顺利完成全部测试。

(3)能自动测试数据并进行分析处理。

对各种形式的测试结果,软件根据其误差范围实时地显示结果,提示检测项目是否符合技术要求,测试点性能好坏一目了然。

5 结束语

以往的外场检查仪多采用面板式开关操作,操作步骤繁琐,自动化程度低,结果记录困难,针对这些问题设计的航空瞄准显示设备微机检测系统,具有检测精度高、操作简单、结果输出方便等优点,且体积小、重量轻,由于全套系统所有部分集成在一个标准手提机箱内,因此便于携带或车载以实现对机载设备的原位检测。使用该系统维护机载设备时,机务人员可根据具体情况,只将检测有问题的部件从飞机上拆下进行维修,正常的部件则可继续使用,这就大大地提高了维护效率和减少了人为差错。同时也为其他机载系统或电子设备的检测提供了一种行之有效的办法。

参考文献:

[1] 孙传友,孙晓斌.测控系统原理与设计[M].北京:北京航

空航天大学出版社,2002.

- [2] 杨成,查光东.基于层次模型的航空火控系统检测数据库系统[J].计算机测量与控制,2004(2):158-160.
- [3] 王少力,吕超.嵌入式计算机模块 PC/104 在工程中的应用[J].光电技术应用,2003(5):73-75.
- [4] 宋辉,蔡忠春.基于 PC/104 的某型飞机发动机控制检测系统的设计[J].装备制造技术,2010(6):55-58.
- [5] 陆庆峰,毛羽刚,黎林坡,等.嵌入式无线视频监控系统的设计与实现[J].计算机技术与发展,2010,20(8):12-15.
- [6] Gue Weina, Deng Hong. Design of Interact Remote Control Hanging Keyboard for Computer Teaching[J]. Control & Automation, 2005(3):1-15.
- [7] 肖忠祥,孟开元.数据采集原理[M].西安:西北工业大学出版社,2001.
- [8] Fedman P, Jennings R. 即学即用 VISUAL C[M].江峡,译.北京:电子工业出版社,1996.
- [9] PC/104 Embedded Consortium. PC/104 Specification Version 2.5[S]. 2003.
- [10] Wolf W. 嵌入式计算系统设计原理[M].孙玉芳,译.北京:机械工业出版社,2002.
- [11] 宋亮,原亮,满梦华,等.军用嵌入式系统中 PCB 设计与测试规范研究[J].计算机技术与发展,2010,20(1):235-239.
- [12] Teng Yuntan, Zhang Lian. Technique Development of PC104 Embedded Module and Its Application in the Geophysical Instrument Design[J]. Acta Seismologica Sinica, 2002(1):107-113.

(上接第 139 页)

挖掘抽取关联规则,帮助进行搭配进货。还有很多地方需要进一步研究,比如:选取规则的标准应该如何设定,商品搭配进货的比例如何计算,进货时间如何确定等都没有涉及;此外只是利用规则进行了最简单的进货规划,没有利用叠加原则进行最终运算,将来会继续对这些问题进行研究。

参考文献:

- [1] 邵峰晶,于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社,2003:93-95.
- [2] Agrawal R, Imielinski T, Swami A. Mining Associations between Sets of Items in Massive Databases[C]//Proc of the ACM-SIGMOD 1993 Int'l Conference on Management of Data. Washington D C: [s. n.], 1993.
- [3] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. [s. l.]: Springer, 1999.
- [4] Houtsma M, Swami A. Set-Oriented Mining for Association Rules in Relational Databases[C]//Proceedings of the 11th IEEE International Conference on Data Engineering. Taipei;

[s. n.], 1995:25-34.

- [5] Agrawal R, Sokant R. Fast Algorithms for Mining Association Rules[C]//Proc of the 20th VLDB Conf. [s. l.]: [s. n.], 1999:33-45.
- [6] 史忠植.知识发现[M].北京:清华大学出版社,2002:22-23.
- [7] 胡纯蓉,刘新华,陈世平.基于 WEB 的比价交易代理模式的研究[J].计算机技术与发展,2011,21(4):228-230.
- [8] 涂承胜,陆玉昌. Web 使用挖掘技术研究[J].小型微型计算机系统,2004,25(7):14-18.
- [9] 向坚持,刘相滨,徐选华.基于用户行为的 Web 使用挖掘数据采集技术研究[J].计算机与现代化,2007(12):59-62.
- [10] 朱志国,邓贵仕. Web 使用挖掘技术的分析与研究[J].计算机应用研究,2008,25(1):29-32.
- [11] 刘立军,周军,梅红岩. web 使用挖掘的数据预处理[J].计算机科学,2007,34(5):200-201.
- [12] 高峰,谢剑英.发现关联规则的增量式更新算法[J].计算机工程,2000,26(12):49-50.