面向零售业的关联规则挖掘的研究与实现

张 珏[1,2],陈 莉[2],田建学[1]
(1. 榆林学院 信息工程学院,陕西 榆林 719000;
2. 西北大学 信息科学与技术学院,陕西 西安 710000)

摘 要:随着零售业在城市的快速发展,智能系统积累了大量的零售业原始数据,急需一种技术来发现数据中蕴含的内在规则,为企业管理者提供决策支持。数据挖掘是目前一个重要的研究方向,可以把日常业务数据知识化。介绍了零售业商务智能系统的发展现状,并通过分析零售业数据来掌握顾客的购买偏好,并同时对挖掘结果进行说明,在一定程度上利用关联规则技术解决现实中的商业问题。针对数量和利润的因素,提出利用频繁项目集寻找商品利润最大化的销售组合模型,零售商可以根据该模型输出的销售组合模型对商品进行捆绑销售,以获得最大利润。提出来竞争商品的概念,即找出隐含在数据库中相互竞争商品的模型,这样就得到了零售业商品推荐模型。实验结果表明,提出的模型能找出高交叉销售利润的商品,在零售业中有很好的实用性。

关键词:数据挖掘;关联规则;零售业商务智能系统;Apriori 算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2016)10-0146-05
doi:10.3969/j.issn.1673-629X.2016.10.032

Research and Realization of Association Rules Mining in Supermarket

ZHANG Jue[1,2], CHEN Li[2], TIAN Jian-xue[1]
(1. Department of Information Engineering, Yulin College, Yulin 719000, China;
2. School of Information Science and Technology, Northwest University, Xi'an 710000, China)

Abstract: With the rapid development of supermarket, a lot of business data are accumulated by intelligent system. It's imperative and necessary to find an effective technique to explore and discover the potential knowledge from the enormous amount of data, which is helpful for business decision making. Data mining has an important research role in the world. It can be used to acquire the knowledge. The current situation of supermarket development is analyzed, and the customer's buying behavior is understood through the analysis of the retail sales data, making the explanation to the mining result, application of association rules to solve real business problems. According to the factors of the quantity and profit, the frequent item sets are adopted to find the sales combination model of profit maximization of commodity, and retailers can use it to bundling and gain the biggest profit. Based on the concept of competitive products, a model is proposed that can be used to find out the hidden in the retail database by the frequent and non-frequent items, getting the model of retail commodity recommendation. The experiment shows that the model can find out the high cross selling goods with good practicality in supermarkets.

Key words: data mining; association rules; retail business intelligence system; Apriori algorithm

0 引 言

随着零售行业竞争的日益激烈,传统的销售模式已经不能适应当前的形势。超市作为日常生活中一种非常频繁重要的零售购物方式,是目前商家争夺的焦点。超市的经营特征决定了超市的特点:雇佣人员少、商品流量大、资金周转快等。现代社会市场变化快,顾客需求也渐渐趋向于个性化,零售业的竞争日趋激烈,超市要获得更多的利润、取得更好的销售业绩就需要最大限度地利用数据库中的大量业务数据。这些数据对企业管理者来说意义重大、价值非凡,通过数理模式分析营销过程中产生的大量数据,划分出不同类型的客户或市场,并分析消费者的偏好和行为,帮助商家保留老客户,发展新客户,提升客户的满意度。比如市场部需要实时获得近两年的销售记录,来向有潜在购买

收稿日期:2015-08-12 **修回日期**:2015-12-24 **网络出版时间**:2016-09-18
基金项目:陕西省自然基金资助项目(2003JM8005);榆林市科技局资助项目(NY13-15);榆林学院青年科技基本资助项目(14YK37)
作者简介:张 珏(1984-),女,讲师,博士研究生,研究方向为大数据、智能信息处理、数据挖掘。

能力、购买意向的客户宣传自己的产品。传统的信息管理系统一直停留在事物处理层面上,只能用于综合统计等一些简单功能,不能获得其他一些较为高级的统计信息,而企业高层决策者则希望通过数据挖掘技术从海量数据中找出更多有用的信息以帮助决策,及时应对市场变化[1-2]。

许多大的零售商由于依据“最佳猜测”来定价而失去了利润,如果他们等待很久才对产品进行打折,或者一些产品本不需要打折,却因为某些不正确的决策对产品进行了打折,则会影响企业的利润。商场的目的是为了追求利润最大化,比如:销售什么样的商品,促销策略如何制定,货架上的商品如何摆放才能吸引顾客眼球,这些都是零售商需要考虑的问题。只有准确掌握顾客的购买偏好才可以帮助商场对类似问题制定准确合理的决策[3-4]。关联规则挖掘就是对商场销售数据进行分析,得到顾客的购买偏好,然后做出合理的决策。在客户消费偏好关联规则分析中,使用数据挖掘中的经典 Apriori 算法对客户消费的商品进行了挖掘,建立合理的客户消费方式模型,根据客户购买偏好进行促销分析,为零售商提供货架摆放和促销方案推荐,更好地把关联规则挖掘用到零售业中,从大量业务数据中找出有用的模式和规律,为客户分析决策服务。

1　关联规则挖掘的相关算法及分析

关联规则挖掘由 R. Agrawal, T. Imielinski 和 A. Swami 提出,应用在数据库上,用来发现零售业中用户购买商品之间的内在的隐含关系及关联规则。关联规则挖掘,就是对业务数据里不同类型的信息进行处理,得到不同类型信息之间的关系,并进一步分析信息之间内在的逻辑规律,为业务运作提供决策支持。决策者还可以利用关联规则提供的信息来合理地设计和安排货架等,从而优化商场布置。

在对商品做促销时,也可以利用关联规则对用户进行分类。例如,英国的某大型超市利用数据挖掘方法对商场上架商品进行关联分析时,发现有一部分滞销商品居然是总计消费额最高的前 25% 顾客的购买对象,于是该商场决定继续销售这批滞销产品,而不是简单地撤下这些滞销产品[5-7]。

基于关联规则的 Apriori 算法有助于挖掘出有用规则,关联规则挖据是在给定的事物数据中找到所有满足最小支持度和最小置信度的形如 $X \rightarrow Y$ 的规则。其中, X 和 Y 分别表示属性集合(称为项集),并且满足 $X \cap Y = \varnothing$。蕴含式 $X \Rightarrow Y$ 称为关联规则, X 、Y 分别称为关联规则的前提和结论。假设 T 为要进行处理的所有事务记录的集合, X 为 T 中包含 X 的事务记录的个数, $X \cap Y$ 为 T 中同时包含 X 和 Y 的有共同属性的事务记录个数, $X \cup Y$ 为 T 中所有事物记录的个数[8-9]。支持度和置信度为:

$$\text{support}(X \Rightarrow Y) = \text{support}(X \cup Y)$$

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{support}(x \cup y)}{\text{support}(x)}$$

支持度衡量整体数据集合的重要性;置信度是描述规则成立的可信度。一般情况下,决策者和用户感兴趣的、有用的关联规则是支持度和置信度都高的关联规则。在统计意义上最小支持度表示项目集合最低重要度,最小置信度则表示规则最低的可靠度。关联规则挖掘问题就是找到支持度和置信度都大于指定阈值的关联规则。如果规则的置信度和支持度都大于最小支持度、最小置信度时,规则是有效的,也称为强关联规则。当数据项集合 X 的支持度大于最小支持度时,就把 X 称为频繁项目集合。

关联规则挖掘算法中涉及的问题主要有两个:如何减少 I/O 操作,因为频繁的 I/O 操作会影响挖掘效率;如何降低需要计算支持率的项目集数量,尽量与频繁项目集的数量接近。

Apriori 算法非常经典,主要分为两大步骤:

第一步骤:找出高频项目集合。

(1)找出高频项目集 $k-1$,若为空,则停止执行。

(2)由(1)中找出任意两个有($k-2$)项目相同的项目集 $k-1$,组合成项目集 k。

(3)判断由(2)找出的项目集,其所有包括的项目集 $k-1$ 是否都出现在(1)中,假如成立就保留此项目集 k;否则就删除。

(4)检查由(3)所获得的项目集 k 是否满足最小支持度,如果符合就加入高频项目集 k;否则就删除。

(5)转到(1)继续查找高频项目集 k_1,循环结束,直到无法产生高频项目集。

第二步骤:产生关联规则。

(1)将所有高频 k 项目集(k_1)拆解成 XY , X , YXY。

(2)判断所有的规则是否符合最小信任度,若符合则称为关联规则。

因为第二步的开销远低于第一步,因此关联规则挖掘的性能主要取决于第一步。

关联规则主要分为以下几类:

(1)根据变量类别的不同,关联规则可分为布尔型和数值型。布尔型处理的值是离散的、种类化的,显示了变量之间的关系。数值型处理的是连续的变量,把多维、多层关联规则结合起来对数据进行动态分割或者直接处理。

(2)根据数据的抽象层次,关联规则分为单层关

联和多层关联。在单层的关联规则中,不考虑变量的现实数据是否属于同一层次;但在多层的关联规则中,就需要考虑数据是否属于同一层次、考虑变量的多层性。

(3)根据规则中数据的维数,关联规则可以分为单维和多维。单维的规则考虑数据的一个维度和单个属性的关系,多维的关联规则考虑处理数据的多个维度和多个属性之间的关系。

2 关联规则应用实例

假定事务数据库如表 1 所示。该数据库中有 4 个事务,设定支持度为 2,置信度为 0.5。

表 1 事物表(1)

记录号	购买的商品
A	苹果、猪肉、鸡蛋
B	苹果、橘子、火腿
C	苹果、橘子、火腿、葡萄
D	橘子、火腿

频繁项集发现过程:

(1)扫描数据集中的所有事务,对每个项的出现次数计数。

(2)最小事务支持度为 2,确定项目集合为 L_1。

(3)为发现频繁项集 L_2,算法连接 L_1 产生候选 2 项集的集合 C_2 = {{苹果,橘子},{苹果,火腿},{苹果,葡萄},{橘子,火腿},{橘子,葡萄},{火腿,尿布}}。

(4)扫描 D 中所有事务,计算 C_2 中每个候选项集的支持数目,如果某个事务包含该候选项集,那么候选项集的支持计数加 1。

(5)确定频繁 2 项集的集合 L_2,它是由 C_2 中大于并等于支持度计数的 2 项集组成,L_2 = {{苹果,橘子},{苹果,火腿},{苹果,葡萄},{橘子,火腿}}。

(6)候选 3-项集的集合 C_3 由频繁 2-项集产生,C_3 = {苹果,橘子,火腿}。

(7)扫描 D 中事务,计算 C_3 中候选集的支持计数,所以{苹果,橘子,火腿}为频繁 3-项集 L_3。

由于频繁项集已经满足最小支持度的要求,所以这时只考虑置信度。设最小置信度为 80%,由频繁项集{苹果,橘子,火腿}可生成强关联规则{苹果,橘子}→{火腿}和{苹果,火腿}→{橘子}。置信度规则如表 2 所示[10-11]。

置信度为 100% 说明,购买前一物品的同时,也会购买后一物品。也就是说购买苹果、橘子的用户同时会购买火腿。购买苹果、火腿的用户同时也会购买橘子。

已知另一数据库数据如表 3 所示。

表 2 置信规则表

关联规则	置信度
{苹果,橘子}→{火腿}	2/2=100%
{苹果,火腿}→{橘子}	2/2=100%
{橘子,火腿}→{苹果}	2/3=67%
{苹果}→{橘子,火腿}	2/3=67%
{橘子}→{苹果,火腿}	2/3=67%
{火腿}→{苹果,橘子}	2/3=67%

表 3 事物表(2)

记录号	购买的商品
A	帽子、围巾
B	帽子、围巾、手套
C	棉衣、手套
D	毛巾、电热毯
E	棉衣、帽子、围巾、手套
F	鞋刷、香皂、空气清新剂
G	浴衣、香皂、剃须泡沫
H	鞋刷、浴衣、香皂、剃须泡沫
I	浴衣、剃须泡沫

数据挖掘过程如下:

第一步:根据定义,计算每种商品的关联规则的支持度。

Support(帽子)= 3/9 = 33%

Support(围巾)= 3/9 = 33%

Support(手套)= 3/9 = 33%

Support(棉衣)= 2/9 = 22%

Support(毛巾)= 1/9 = 11%

Support(电热毯)= 1/9 = 11%

Support(鞋刷)= 2/9 = 22%

Support(香皂)= 3/9 = 33%

Support(浴衣)= 3/9 = 33%

Support(剃须泡沫)= 3/9 = 33%

Support(空气清新剂)= 1/9 = 11%

第二步:设定最小的支持度阈值为 20%,将大于或等于最小支持度阈值的商品挑选出来,那么帽子、围巾、手套、棉衣、鞋刷、香皂、浴衣、剃须泡沫可以被挑选出来。

第三步:计算商品关联规则的置信度。

帽子、围巾、手套、棉衣、鞋刷、香皂、浴衣和剃须泡沫的置信度为:

Confidence(棉衣 ⇒ 帽子)= Support(棉衣 ⇒ 帽子)/Support(棉衣)= 11%/22% = 0.5

Confidence(棉衣 ⇒ 围巾)= Support(棉衣 ⇒ 围巾)/Support(棉衣)= 11%/22% = 0.5

Confidence(棉衣 ⇒ 手套)= Support(棉衣 ⇒ 手

套)/Support(棉衣)= 22%/22% =1

Confidence(毛巾 ⇒ 毛毯)= Support(毛巾 ⇒ 毛毯)/Support(毛巾)= 11%/11% =1

Confidence(香皂 ⇒ 浴衣)= Support(香皂 ⇒ 浴衣)/Support(香皂)= 22%/33% =0.67

Confidence(香皂 ⇒ 剃须泡沫)= Support(香皂 ⇒ 剃须泡沫)/Support(香皂)= 22%/33% =0.67

Confidence(浴衣 ⇒ 剃须泡沫)= Support(浴衣 ⇒ 剃须泡沫)/Support(浴衣)= 33%/33% =1

Confidence(鞋刷 ⇒ 剃须泡沫)= Support(鞋刷 ⇒ 剃须泡沫)/Support(鞋刷)= 11%/22% =0.5

其余数据通过 $X \Rightarrow Y$ 的信任度表示如表 4、5 所示。

表 4　置信度(1)

商品	棉衣	帽子	围巾	手套	电热毯	毛巾
棉衣	1	0.5	0.5	1	0	0
帽子	0.33	1	1	0.67	0	0
围巾	0.33	1	1	0.67	0	0
手套	0.67	0.67	0.67	1	0	0
电热毯	0	0	0	0	1	0
毛巾	0	0	0	0	1	1

表 5　置信度(2)

商品	鞋刷	香皂	空气清新剂	浴衣	剃须泡沫
鞋刷	1	1	0.5	0	0.5
香皂	0.67	1	0.5	0.67	0.67
空气清新剂	1	1	1	0	0
浴衣	0.33	0.67	0	1	1
剃须泡沫	0.33	0.67	0	1	1

第四步:设定最小信任度阈值为 0.6,得到的规则如表 6 所示。

根据上述生成的关联规则,可以发现顾客潜在的购买习惯和偏好,将帽子、围巾放置在一起,以方便顾客选购,甚至可以把帽子、围巾和手套,棉衣和手套放在一起捆绑销售。浴衣和剃须泡沫可以捆绑在一起销售,香皂、鞋刷和剃须泡沫也可以在一起捆绑销售。在进货的时候,可以考虑将上述商品统一采购,也可以放在一起统一印发促销广告,来提高商品的支持度和信任度。上述关联规则生成中,任务度和支持度都高的就可以考虑在一起捆绑销售,让消费者交叉购买以提高消费力[12-13]。

大型零售业的利润主要来自于以下三个方面:商品差价、供应链成本和管理成本。但在目前激烈竞争的客观环境下,大型零售业在以上三方面利润上升空间很小。想要在市场中拥有竞争力关键在于提高销售额,即吸引更多的顾客,并要提高顾客的购买金额,这样零售业才能获得更多更高的利润、在市场中拥有更

表 6　规则表

$X \Rightarrow Y$ 规则	支持度	信任度
棉衣 ⇒ 手套	22%	1
帽子 ⇒ 围巾	33%	1
帽子 ⇒ 手套	22%	0.67
围巾 ⇒ 帽子	33%	1
围巾 ⇒ 手套	22%	0.67
手套 ⇒ 棉衣	22%	0.67
手套 ⇒ 帽子	22%	0.67
手套 ⇒ 围巾	22%	0.67
毛巾 ⇒ 电热毯	11%	1
鞋刷 ⇒ 香皂	22%	1
香皂 ⇒ 鞋刷	22%	0.67
香皂 ⇒ 浴衣	11%	0.67
香皂 ⇒ 剃须泡沫	22%	0.67
空气清新剂 ⇒ 鞋刷	11%	1
空气清新剂 ⇒ 香皂	11%	1
浴衣 ⇒ 香皂	11%	0.67
浴衣 ⇒ 剃须泡沫	33%	1
剃须泡沫 ⇒ 香皂	22%	0.67
剃须泡沫 ⇒ 浴衣	33%	1

多的竞争力。超市促销就是为了提高营业额,利用各种方法和手段,让消费者能够了解并且注意到超市的产品从而刺激消费者的购买欲望,最终促使消费者实现购买行为,因此促销是零售业日常非常重要的一项工作。以往零售商考虑的角度就是单个商品的利润,但在实际经营中,很多时候最大利润来源于商品组合销售。消费者心理学指出,消费者的购买决策带有很强的情景性,顾客是否购买商品会随着情景的变化而变化,所以零售业货架的安排和设计变得尤其重要,目标就是让顾客发现更多的商品,进而产生购买冲动,那么可以建议在摆放商品时,尽量将置信度较高的商品摆放在和人视线平行的货柜以方便用户购买,同时也能促进相邻商品的销售量上涨。通过统计数据还发现,最大置信度和最小支持度相差越小,生成的竞争商品组数越多。如果两种物品的置信度都较低,说明这两种物品之间在购买时没有关联关系,那么就可以分开摆放。对于这些支持度、置信度较小的商品可以采取一些措施来提高销售业务和顾客满意度,进而提升企业竞争力。一般来说采取以下策略来激发顾客购买的欲望:

(1)制定促销活动,利用关联规则确定不同商品销售的关联关系,来精确确定促销对象。

(2)对销售、顾客、产品、时间和地区进行分类分析,考虑到不同顾客的需求、不同产品的销售和不同品

牌日用品的质量、价格、利润等,对不同维度进行分类,这样就可以更准确地掌握顾客类型、产品畅销程度,以及在不同时间、不同地域的销售区别。

(3)分析顾客购买趋势,对顾客在不同时期购买的商品进行分析,分析顾客消费变化的原因,然后及时调整商品的价格和种类,挽留老顾客,吸引新顾客。

(4)如果某种产品存在它的竞争商品,那么企业可以把购买竞争商品的顾客列为重点顾客,并且商品缺货时,竞争商品就可以作为临时替代品进行销售。

(5)高推荐度的商品个体利润都比较低,超市可以把推荐度高的商品陈列在显眼的地方,这样在节省顾客购买时间的同时也可以增加相关联商品的销售额[14-15]。

(6)进行捆绑销售,比如优惠购买,消费者购买 A 产品时可以用低于正常价格的形式购买到 B。比如统一出售,多种产品按照捆绑后低于单独标价的价格出售,这样在降低了销售成本的同时,也增加了销售额和顾客满意度,起到“1+1>2”的效果,让产品相互协调和促进。

(7)交叉销售,在货架摆放时把关联程度高的商品由过去的就近摆放调整为远离摆放,比如可以交换个人卫浴和厨具餐具的位置,使洗衣用品和卫生清洁品相对远离,这样购买这两类产品的顾客就需要穿过厨具和家居日用品区,这样就可能引起消费者的购物冲动。经常性有意识地改变超市货架布局,来打破消费者的购买习惯,使消费者发现没有注意到的商品,吸引消费者购买,以提高营业额,而同时超市应该将日销售量高的物品摆放在两端,销售量低的产品摆放在容易引人注意的地方,这样消费者在购买的时候可以快速定位,也可以引起顾客对各个货架的关注,从而浏览整个货架,带动更多的销售量。

3 结束语

关联规则技术是一项重要的数据挖掘技术,该技术可以从海量的业务数据中挖掘消费者的购买行为的关联性。关联规则能挖掘不同种类项目之间的相关性,因此,可以找出潜在的商品销售的关联性和客户消费倾向等信息。当然如果考虑序数资料间存在相似度,可以找出更多有意义的规则。然而,在产生更多有意义的高频项目集的同时也会产生相似度太低的项目集。为了解决上述问题,文中以 Apriori 算法为基础,挖掘出具有高度关联性的关联规则,针对挖掘结果提出了包括捆绑销售、竞争分析、交叉营销、商品推荐等不同的解决方案,利用挖掘出的有意义信息,企业可进行决策参考,实现了关联规则挖掘在零售业实体中的应用研究,对零售业的发展有着较为重要的现实意义。

参考文献:

[1] Han J W,Kamber M. 数据挖掘概念与技术[M]. 范 明,孟小峰,译. 北京:机械工业出版社,2006.

[2] 亓文娟,晏 杰. 关联规则挖掘在超市中的应用研究[J]. 吉林师范大学学报:自然科学版,2013,34(2):138-141.

[3] 陈 莉,焦李成. 文档挖掘与降维技术[J]. 西北大学学报:自然科学版,2003,33(3):267-271.

[4] 李颖基,彭 宏,郑启伦,等. Web 日志中有趣关联规则的发现[J]. 计算机研究与发展,2003,40(3):435-439.

[5] 卜耀华. 关联规则挖掘技术在零售业中的应用[J]. 商场现代化,2009(10):97-98.

[6] 张小利,陈 莉. 数据挖掘在智能交通系统中的应用[J]. 西北大学学报:自然科学版,2005,35(6):687-690.

[7] Heravi M J,Zaiane O R. A study on interesting measures for associative classifiers[C]//Proceedings of the 2010 ACM symposium on applied computing. Sierre:ACM,2010:1039-1046.

[8] Ting S L,Tse Y K,Ho G T S,et al. Mining logistics data to assure the quality in a sustainable food supply chain:a case in the red wine industry[J]. International Journal of Production Economics,2013,152:200-209.

[9] Liao Shu-Hsien,Chu Peihui,Hsiao Pei-Yuan. Data mining techniques and applications - a decade review from 2000 to 2011[J]. Expert Systems with Applications,2012,39:11303-11311.

[10] Rong Jia,Vu H Q,Law R,et al. A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining[J]. Tourism Management,2011,33(4):731-740.

[11] Liou J J H,Tzeng G H. A dominance-based rough set approach to customer behavior in the airline market[J]. Information Sciences,2010,180(11):2230-2238.

[12] 王伟辉,耿国华,陈 莉. 数据挖掘技术在保险业务中的应用[J]. 计算机应用与软件,2008,25(3):123-125.

[13] 闫 珍. 面向零售业的关联规则动态挖掘算法研究[D]. 南京:南京航空航天大学,2010.

[14] 黄嘉满. 面向零售业的关联规则挖掘的研究与实现[D]. 上海:上海交通大学,2007.

[15] Han Jiawei, Kamber M. Data mining: concepts and techniques[M]. 2nd ed. Beijing:China Machine Press,2011:146-155.