

加权频繁项集挖掘算法在超市商品捆绑销售中的应用

姜 薇, 张学芹

(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

摘 要: 引入一种新的加权关联规则支持度和置信度的计算方法, 并利用矩阵的存储结构提出一种新的加权关联规则挖掘算法, 从而改进了加权频繁项集的挖掘效率. 该算法在 Apriori 算法的基础上, 对数据库仅需扫描一次, 能很快地计算项集的支持度, 大大减少了 I/O 次数, 有效提高了加权频繁项集的生成效率. 通过应用于超市捆绑销售, 证明了该算法能有效地提取商品间的关联信息, 有助于商品的销售.

关键词: 数据挖掘; 加权关联规则; 频繁项集; 捆绑销售

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2011)10-0142-04

The Application of Weighted Frequent Itemsets Mining Algorithm in Binding Sales of Supermarket

JIANG Wei, ZHANG Xue-qin

(Computer Department, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Introducing a new calculation of weighted support and confidence, this paper proposes a weighted association rule mining algorithm based on matrix, which can improve the weighted frequent itemsets mining methods. Based on Apriori algorithm, it used matrix's storage structure, which scans the database only once and can quickly calculate the supporting degree of itemsets. Therefore, this algorithm greatly reduces the I/O, and can improve the weighted frequent itemsets generation effectively. By being used in the binding sales of supermarkets, shows that the algorithm can extract the relationship between product information and help the sales of goods effectively.

Key words: data mining; weighted association rules; frequent Items; binding sales

1 引言

随着经济的发展, 商品交易的种类以及数量正以惊人的速度在增长, 即时有效的分析客户的购买行为及模式无疑对商家的销售起到推波助澜的作用. 一般来说, 大型连锁超市是^[1]一种利润率相对较低的业态, 利润低的主要原因是大型超市的低价竞争战略, 通过吸引更多的顾客, 提高顾客的光顾次数, 并且提高每个顾客的单次的消费金额, 超市才得以获取更多的利润总额. 超市中每天都会积累大量的数据, 如何能有效地利用数据挖掘知识对这些数据进行处理, 以期提取出有用的规则正成为研究的热点.

关联规则作为数据挖掘中的一项重要技术, 主要用于发现数据中项集之间的关联性. Apriori^[2]算法是关联规则最经典的算法, 它由 Rakesh Agrawal 和 Ramakrishnan Srikant 于 1994 年首次提出. 后来的很多算法研究都是基于对其的改进. 该算法的主要缺点是在产生频繁模式的过程中, 需要产生大量的候选项和多次遍历数据库, 占用大量的内存空间和 CPU 处理时间.

如何既考虑属性数量和属性权值对关联规则的影响, 保持 Apriori 算法的封闭性, 又能快速有效地挖掘出加权关联规则? 为此, 本文提出了一种采用矩阵存储结构的加权关联规则挖掘算法, 并将其应用于超市捆绑销售中.

2 加权关联规则

2.1 关联规则的概念

$I = \{i_1, i_2, \dots, i_m\}$ 是 m 个不同项目的集合. 现给定一个交易数据库 D , D 中的每个元素 T 是 I 中项目组成的集合, T_{id} 可以唯一地标识每个交易.

关联规则是型如 $X \Rightarrow Y$ 的蕴含式, 其中 $X \subset I, Y \subset I$, 且 $X \cap Y = \emptyset$.

关联规则分析又称购物篮分析, 是从庞大的交易记录数据库中寻找出商品项目之间的关联性. 主要是探讨在大型数据库中, 商品项目与其属性之间共同发生的关联, 以此来分析消费者的购买行为及购买商品的消费模式.

挖掘过程如下:

(1) 找出所有频繁项集.

(2) 通过与最小支持度 $\min_sup$ 和最小置信度 $\min_conf$ 的比较, 由频繁项集生成强关联规则.

由于强关联规则的生成相对来说比较容易, 因此研究的重点在于如何快速发现所有频繁项集.

2.2 Apriori 算法

Apriori 算法是关联规则最重要的算法之一, 它以概率为基础, 能够循序渐进的找出数据库中各项集之间的关系, 以此来形成规则.

设候选 k -项集为 C_k , C_k 中频率大于或等于最小支持数的 k -项集为 L_k .

该算法的基本过程如下:

(1) 计算所有的 C_1 ;

(2) 扫描数据库, 删除非频繁子集, 生成 L_1 ;

(3) L_1 与自己连接生成 C_2 ;

(4) 扫描数据库, 删除 C_2 中的非频繁集, 生成 L_2 ;

(5) 依次类推, 通过 L_{k-1} 与自己连接生成 C_k , 再次扫描数据库, 生成 L_k , 直到生成所有的频繁项集.

产生频繁项集的基本思想由连接和剪枝两步完成:

(1) 连接: 通过 L_{k-1} 与自身连接产生候选 k -项集. 候选项集记做 C_k .

(2) 剪枝: 所有的频繁 k -项集都包含在 C_k 中. 扫描数据库, 确定 C_k 中每个候选集计数, 从而确定 L_k .

2.3 Apriori 算法的主要缺陷

Apriori 算法存在两个主要缺陷:

(1) 需要产生大量候选项集

第一频繁集两两合并产生第二候选集, 因此若第一频繁项目集有 k 个项目, 则第二候选集将有 $(k-1) + (k-2) + \dots + 1$ 个, 这个数量将是非常巨大的.

(2) 需要重复扫描数据库以进行计数

由于每一个候选项目集都要重新扫描一遍数据库, 再求其支持度, 因此执行效率较低.

自从 Apriori 算法提出以后, 大量的研究工作都是针对如何减少扫描次数以及如何减少读取数据库的 I/O 次数进行的.

2.4 加权关联规则概述

传统关联规则挖掘只考虑分析的项在数据库中出现的频率, 不考虑项在记录中的数量, 并认为数据库中各项目的重要性相同. 然而在现实世界中, 各项目的重要性通常不尽相同, 于是加权关联规则被提出, 并得到了越来越多的关注与研究.

文献[3]中在 Apriori 基础上提出加权关联规则概念, 并给出两种加权关联规则的挖掘算法: MINWAL(O) 和 MINWAL(W). 它引入权重的概念, 但存在计算过程中支持度可能大于 1 的隐患. 文献[4]中考虑属性数量和属性权值对关联规则的影响, 提出一种新的加权支持度和加权置信度计算方法, 通过改进加权支持度设置模型, 保持了 Apriori 算法的向下封闭性. 本文结合了前述文献的优点, 提出了适合于加权关联规则挖掘的一种新算法.

3 加权关联规则模型

3.1 超市商品数据的特性

一般来说, 超市可以通过压缩管理成本、优化供应链以及商品进销来产生更大的利润. 由于这几个方面都没有太大的利润空间, 所以超市可以依靠自身的促销计划为提高自身的竞争优势. 如通过将相关联的商品摆放在比较近的货柜, 如经典的“啤酒与尿布”, 这样在方便顾客的同时还能大大促进不同商品的销售量.

与传统的数据挖掘相比, 商家对获利高的商品可能有更高的关注度, 而对获利较低商品的重视程度要低一些, 这就可以通过设定不同的权值来进行挖掘, 以做到不同数据不同对待.

3.2 加权支持度与加权置信度

定义 1 项目权值是项目集为每一个项目赋予 $[0,1]$ 之间的一个非负实数作为权值,以体现该项目的重要度,记为 $i_k[w]$.

定义 2 $I = \{i_1, i_2, i_3, \dots, i_k\}$ 是由 k 个不同项目组成的集合, $D = \{T_1, T_2, T_3, \dots, T_n\}$ 是事务数据库,其中,每个事务 $T_i (i = 1, 2, \dots, n)$ 包含事务唯一标识 T_{id} 和一个 I 中项的子集 X .

文献[4]中考虑属性数量和属性权值对关联规则的影响,对第 i 个事务中的第 j 个项目的加权值记为 $t_i[i_j[w]]$.

定义 3 项集事务权值是一个数据库中项目集中各项目的权值汇总,项目集 x 在事务 t 中的项目权值计算为

$$ITW(X) = \frac{\prod_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}{\sum_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}$$

定义 4 加权支持度是事务数据库中包含该项目的事务项集权值的汇总.

$$WS(X) = \frac{N_x \prod_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}{n \sum_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}$$

定义 5 加权置信度是事务数据库中的加权支持度与 $X \cup Y$ 加权支持度的比值: $WC(X \rightarrow Y) = \frac{WS(X \cup Y)}{WS(X)}$

3.3 矩阵存储结构

定义 6 1-项集 I_j 的向量定义为

$$D_j = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \dots \\ t_{nj} \end{bmatrix}, \text{其中}, t_{ij} = \begin{cases} 0, & I_j \notin T_i \\ 1, & I_j \in T_i, \end{cases}$$

支持度计数 $\text{support_count}(I_j) = \sum_{i=1}^n t_{ij}$.

定义 7 2-项集的向量定义为

$$D_{ij} = D_i \wedge D_j = \begin{bmatrix} d_{1i} \wedge d_{1j} \\ d_{2i} \wedge d_{2j} \\ \dots \\ d_{ni} \wedge d_{nj} \end{bmatrix},$$

$\text{support_count}\{I_i, I_j\} = \sum_{k=1}^n (d_{ki} \wedge d_{kj})$.

定义 8 k -项集 $\{I_1, I_2, \dots, I_k\}$ 的向量定义为

$$D_{12} \dots_k = D_1 \wedge D_2 \wedge \dots \wedge D_k = (D_1 \wedge D_2 \wedge \dots \wedge D_{k-1}) \wedge D_k,$$

$$\text{support_count}\{I_1, I_2, \dots, I_k\} =$$

$$\sum_{q=1}^n \{(d_{q1} \wedge d_{q2} \wedge \dots \wedge d_{q(k-1)}) \wedge d_{qk}\}.$$

3.4 算法描述

3.4.1 算法思想

由于引入的新的加权支持度、置信度的计算方法能保持频繁集的向下封闭性,因此算法是满足向下封闭的,证明如下.

证明:频繁集的子集也是频繁集.

$I = (i_1, i_2, \dots, i_m)$ 是 m 个不同项目的集合. 交易数据库 $D = \{T_1, T_2, T_3, \dots, T_n\}$ 是事务数据库, D 中的每个元素 T 是 I 中项目组成的集合, $T \subseteq I$ 最小支持度为 t . 设 Z 为频繁集, $Z = X \cup i_j$, i_j 是 I 的一项,显然 $X \subset Z$.

$$\because WS(X) = \frac{N_Z \prod_{k=1}^{|Z|} (\forall [i[w] \in Z] t_i[i_k[w]])}{n \sum_{k=1}^{|Z|} (\forall [i[w] \in Z] t_i[i_k[w]])} =$$

$$\frac{N_Z(i_j[w] \prod_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}{n(\sum_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]]) + i_j[w])} > t,$$

$$\text{又} \because 0 \leq i_j[w] \leq 1, N_Z \leq N_X,$$

$$\therefore WS(X) =$$

$$\frac{N_X \prod_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])}{n \sum_{k=1}^{|X|} (\forall [i[w] \in X] t_i[i_k[w]])} \geq WS(Z) > t.$$

$\therefore X$ 也是频繁的,得以证明.

由此可以利用 Apriori 算法的思想进行加权关联规则的挖掘:先扫描一遍数据库,将所有项集存储在矩阵上,然后读取矩阵信息计算加权支持度和置信度,生成频繁项集. 仅扫描一次数据库,就可实现加权关联规则的挖掘.

3.4.2 算法步骤

(1) 根据用户需要,为每个项目分配权值,制定最小加权支持度和最小加权置信度.

(2) 扫描数据库,按照定义 6 将相关数据存入矩阵.

(3) 根据矩阵存储信息计算支持度生成频繁 1-项集.

(4) 按照定义 7 生成频繁 2 项集.

(5) 按照 Apriori 的连接条件由 $(k-1)$ -项集生成 k -项集. 利用支持度计数计算加权支持度,

将计算结果与最小加权支持度进行比较:若小于最小加权支持度,则将其删除;否则,计入频繁 K 项集的候选集.

(6) 重复上述步骤,与最小加权支持度比较,产生更高层次的频繁项的集合,再结合产生下一级候选项集,直到不再产生新的候选项集为止.

4 算法在超市商品捆绑销售中的应用

以某超市数据为例,应用算法说明加权关联规则在商品捆绑销售中的应用.

商品分别为: A(洗衣皂), B(牙膏), C(酱油), D(食用油), E(抽纸)

事务数据库如表 1 所示.

表 1 事务数据库

T_{id}	A	B	C	D	E
1	1	0	1	1	0
2	1	0	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1

依次为每种商品 ABCDE 分配权值为: 0. 6, 0. 1, 0. 5, 0. 9, 0. 8,取最小加权支持度 $Wmin-sup=6\%$

计算得到频繁 1 项集 $L_1 = \{A,C,D,E\}$.

频繁 2 项集 $L_2 = \{AC,AD,AE,CD,CE\}$.

频繁 3 项集 $L_3 = \{ACE\}$,即{洗衣皂,酱油,抽纸},算法结束.

与传统关联规则相比,此算法挖掘出的数据更能满足商家的需要,而且更加迅速高效.

5 结束语

本文研究加权关联规则挖掘问题,根据超市

商品交易数据的特点,提出基于矩阵的加权频繁项集生成算法. 由于加权支持度和置信度的计算方法保证了向下封闭性,通过存储在矩阵中的数据,方便的计算出项集的支持度. 仅需一次扫描数据库,大大提高了算法的效率.

目前,数据挖掘技术在我国的应用还不太普遍,今后研究的重点将结合大型超市多维性、动态性、数据库超大规模性的特点,以期将研究成果应用于实际中.

参考文献:

[1] 于芳,马维忠. 关联分析在超市商品捆绑销售中的应用[J]. 商场现代化,2010, 599(1) : 33-34.
[2] Agrawal R, Imielinski T, Wami A S. Mining association rules between sets of items in large databases[C] // Proc of the ACM SIGMOD Conference on Management of Data, Washington: ACM,1993(05): 207-216.
[3] Cai Chun Hing, Fu Ada Waichee, Cheng Chunhung, et al. Mining association rules with weighted items [C]//Proc of International Database Engineering and Applications Symposium. Cardiff, Wales, UK: [s. n], 1998.
[4]李成军,杨天奇. 一种改进的加权关联规则挖掘方法[J]. 计算机工程,2010,36(7):55-57.

作者简介:

姜 薇 女,(1966-),副教授,硕士研究生导师. 研究方向为数据挖掘;
张学芹 女,(1977-),硕士研究生. 研究方向为数据挖掘.

(上接第 141 页)

1st CJK Joint Workshop on Pattern Recognition, CJK-PR. 2009; 349-353.
[13] 党向盈, 吴锡生,赵勇. 新的基于边缘检测提高图像质量的插值算法[J]. 计算机工程与设计, 2007, 28(15): 3636-3639.
[14] 章毓晋, 图象工程(下册)—图象理解与计算机视觉[M], 北京: 清华大学出版社. 2000.

作者简介:

严云洋 男,(1967-),博士,教授. 研究方向为数字图像处理、模式识别;
郭志波 男,(1975-),博士,副教授. 研究方向为模式识别,数字图像处理;
高尚兵 男,(1981-),博士研究生,讲师. 研究方向为计算机视觉、模式识别;
唐岩岩 女,(1986-),硕士研究生. 研究方向为数字图像处理、模式识别.