基于关联规则数据挖掘 Apriori 算法的研究与应用

郭 涛,张代远
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:目前在我国,对数据挖掘技术的研究与应用并不是很广泛。大多数数据库只能实现数据的录入、查询、统计等较低层次的功能,无法发现数据中存在的各种有用的信息。基于关联规则的数据挖掘主要用于发现数据集中项目之间的联系。以超市购物为例,目的在于找出顾客所购买商品之间的内在关联。利用 Apriori 算法的先验原理,减少 Apriori 算法在搜索频繁项目集时对候选式的搜索次数,并在对顾客购买的商品模型进行抽象的基础上,利用 vc++与 access 数据库实现的算法系统,对所购买的商品之间的内在关联进行模拟分析。根据得到的数据分析出置信度较高的几种商品,通过对这些商品集中摆放,可以提高收益,从而证明改进的 Apriori 的实用性。

关键词:数据挖掘;关联规则;Apriori 算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2011)06-0101-03

Research and Application on Association Rules Based on Apriori Algorithm

GUO Tao, ZHANG Dai-yuan
(Coll. of Computer, Nanjing Univ. of Posts and Telecommunications, Nanjing 210003, China)

Abstract: At present in China, data mining research and application is not widely used. Most of the database only for data entry, query, statistics and other lower-level functions, can not find the data that exists in a variety of useful information. Association rules found in data mining is mainly used for the relevant links between data items. Supermarket shopping is taken for example, to find relevancy of customers' buying. Applying priori principle of Apriori, the number of searching for frequent item sets is reduced. On the foundation of abstracting model of customers' buying and implementation of algorithm-based system based on vc++ and access database, intrinsic correlation of goods purchased is simulated and analyzed. Income will be increased, if a few commodities with a high degree of confidence are put together. According to above-mentioned theory and analysis, the practicality of improved-algorithm has been proved.

Key words: data mining; association rules; Apriori algorithm

0 引 言

关联规则最早运用于超市的购物篮,关联规则概念提出的目的在于揭示给定数据集中数据项之间内在关联以及存在的各种有用的信息,根据所挖掘的潜在的依赖关系,可以从一个数据项的信息来推断其他相关联的数据项的信息[1]。如今关联规则已经被推广到许多领域,只要涉及到从大型的数据集中获取知识的问题,关联规则都可能成为有力的工具。文中通过对 Apriori 算法的研究,设计实现了一个关联规则挖掘算法的原型系统,对某超市在一个月内的顾客购买商品情况进行抽样数据处理,得出相关结果并对其进行分析。

1 数据挖掘与关联规则

1.1 数据挖掘

当前,数据挖掘公认的定义是由 Fayyad 给出的:数据挖掘是一个用以确定数据中有效的、未知的、新颖的、具有潜在可用性并且最终可被理解的模式的重要处理过程[2,3]。

1.2 关联规则

关联规则是指在日志数据、关系数据或者其他信息载体中,存在于项目集合或对象集合之间的频繁模式、相关性或因果结构。关联规则的获取主要是通过数据挖掘的方法从大量的事件记录数据库中找出那些频繁模式[4]。

收稿日期:2010-11-25;**修回日期**:2011-03-03
作者简介:郭 涛(1987-),男,硕士,研究方向为智能计算、神经网络;张代远,教授,研究方向为神经网络、演化计算、计算机体系结构。

关联规则的传统算法步骤是：首先找出所有的频繁项目集，然后由频繁项目集产生满足最小置信度和最小支持度的规则。关联规则中的支持度和置信度分别用来衡量规则的有效性和可信度。若存在规则 $X -> Y$，则该规则的支持度表示事务集合中包含 $X \cup Y$ 中的所有项目的事务的出现频度。支持度是一个有效的评价指标，如果支持度的值太小，就表明相应的规则在整个事务集合中只是偶然出现，在商业应用中，该规则很可能没有价值。对于置信度而言，若存在规则 $X -> Y$，则该规则的置信度表示 Y 在包含 X 的事务中出现的频繁程度。置信度的大小决定了规则的可预测度的大小。如果所选规则的置信度值太小，就表明从 X 就很难可靠地推断出 Y。同样，置信度太低的规则也很可能没有价值[5,6]。

基于关联规则的算法以 Apriori 算法为代表，其后的 MPL 等算法大多是在 Apriori 算法的基础上衍生或者改进[7]。

2 Apriori 算法分析

2.1 算法基本思想

Apriori 算法基本思想：首先，计算含有一个元素的项目集出现的频率，找出那些不小于最小支持度的项目集，得到一维最大项目集，生成一维频繁集。然后进行连接运算生成二维候选集，再根据预先给定的最小支持度，生成二维频繁集。重复上述过程，直到生成 M 维频繁集，并且不能再生成满足最小支持度的 $(M+1)$ 项目集。这里有一点需要注意：若存在 K 维候选集 $(K=3,\cdots,M)$，其中某个元素的 $(K-1)$ 子集不是 $(K-1)$ 维频繁集，则该候选集将被删除[8]。

Apriori 算法的流程如图 1 所示。

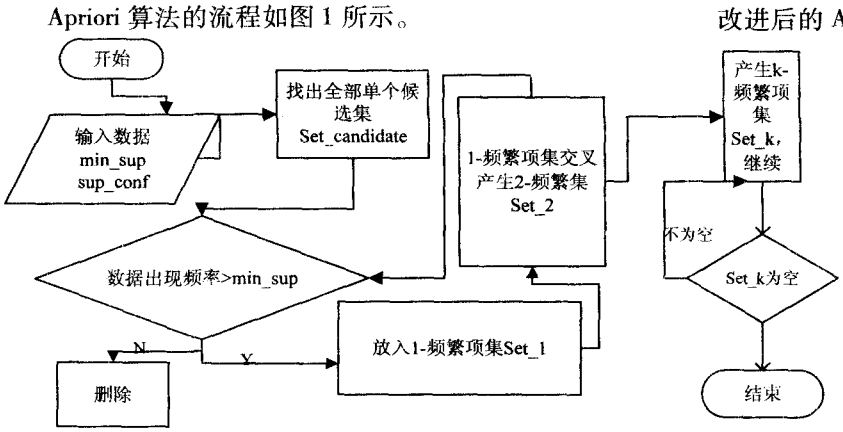

图 1 Apriori 算法流程图

2.2 Apriori 算法的缺陷

Apriori 算法存在的缺陷，由上述 Apriori 算法的基本思想可知：在每一次计算候选项目集时会循环产生过多的项目组合，其中包含许多不应该参与组合的元素，没有进行预处理。除此之外，在 K 维候选集，判断某个元素的 $(K-1)$ 子集不是 $(K-1)$ 维频繁集时，需要在数据库进行全部的比较[9]，这种不断的重复匹配比较，会增加系统的 I/O 负担[10]。

2.3 算法的改进

首先在这里给出 Apriori 算法的先验原理：

如果一个项目集是频繁的，那么它的所有子集都是频繁的。先验原理成立的原因：

$$\forall X,Y:(X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

即一个项目集的支持度不会超过其任何子集的支持度，该性质称作支持度的反单调性质。Apriori 先验原理主要用于搜索频繁项目集时对候选式筛选的过程，利用先验原理能够比较好地避免盲目的搜索，提高频繁项集的查找效率（见图 2）。

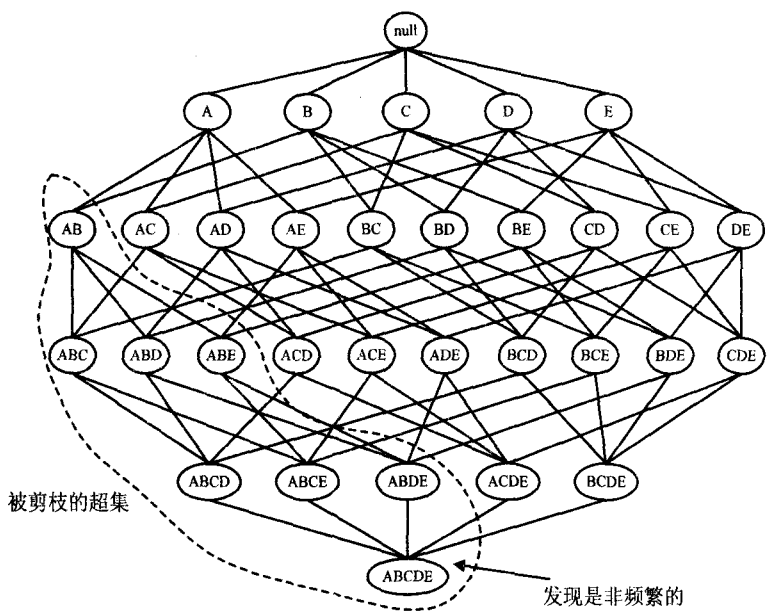

图 2 改进的 Apriori 算法

改进后的 Apriori 算法流程[11,12]如下：

设定 $k = 1$

扫描事务数据库一次，生成频繁的 1 - 项集

如果存在两个或以上频繁 k - 项集，重复下面过程：

[候选产生] 由长度为 k 的频繁项集生成长度为 $k+1$ 的候选项集

[候选前剪枝] 对每个候选项集，若其具有非频繁的长度为 k 的子集，则删除该候选项集

[支持度计算] 扫描事务数据库一次，统计每个余下的候选项集的支持度

[候选后剪枝] 删除非频繁的候选项集，仅保留频繁的 $(k+1)$ - 项集

设定 $k = k + 1$

其中改进后的 Apriori 算法的核心步骤如下:

候选产生:

设 $A = \{a_1, a_2 \cdots a_k\}$ 和 $B = \{b_1, b_2 \cdots b_k\}$ 是一对频繁 k - 项集,当且仅当 $a_i = b_i (i = 1, 2 \cdots k-1)$ 并且 $a_k \neq b_k$ 时,合并 A 和 B,得到 $\{a_1, a_2 \cdots a_k, b_k\}$。比如合并{Bread, Milk}和{Bread, Diaper}得到{Bread, Milk, Diaper},但{Milk, Bread}和{Bread, Diaper}不能合并。

候选前剪枝:

设 $A = \{a_1, a_2 \cdots a_k, a_{k+1}\}$ 是一个候选 $(k+1)$ - 项集,检查每个 A' 是否在第 k 层频繁项集中出现,其中 A' 由 A 去掉 $a_i (i = 1, \cdots, k-1)$ 得到,若某个 A' 没有出现,则 A 是非频繁的。

3　基于 Apriori 算法的应用

3.1　测试数据模型

文中所用系统利用 Apriori 的改进算法,采用 VC++语言结合 ACCESS 数据库编写而成。

文中提供 15 种测试商品,每个商品用小写英文字母表示,如从 a 到 o 进行编号。

整理出某个超市在一个星期内的销售数据,每天固定 100 个客户,星期六星期天顾客比平时多,则这两天,每天抽出 250 个顾客购物信息。

在试验中,Apriori 算法计算每天的前 100 客户采用 min_sup = 0.1, min_conf = 0.45 或 min_conf = 0.55,其他全部采用 min_sup = 0.1, min_conf = 0.45。

下面用一个例子说明对所采集的数据预处理的步骤,如一个客户发票打印的数据为:

日期:2010-11-20

代号	商品名称	数量	价格(元)
11042	康师傅红烧牛肉面	1	1.80
11043	康师傅鲜虾鱼板面	1	1.80
12153	水晶阿胶枣	1	3.60
15412	光明高钙牛奶	12	27.6

忽略商品的代号、数量、价格以及购买时间。对要进行测试的顾客从 0 开始进行编号。不同的商品名称、商品类型用不同的英文字母表示。在进行测试数据预处理的时候,根据不同的实际需求,划分各种商品的类型。如上述超市发票单上的康师傅红烧牛肉面和康师傅鲜虾鱼板面都算成方便面类型。但如果是想计算方便面之间的关联关系,则可把两种方便划分成不同的分类。文中把方便面都划为一类,用字母"a"表示。上述例子可以处理成以下的形式,其中 b、c 分别表示水晶阿胶枣和光明高钙牛奶:

顾客编号	购买的商品的代号
0	a b c

数据经过预处理后,以 txt 文本的形式存放,通过文件输入流的形式输入到系统中,再进行相关的计算。测试数据经过预处理后,一共有 1000 个顾客,留下顾客编号和商品代号。

3.2　测试结果分析

以某个月内 1000 位顾客购买商品的数据测试为例,分析商品之间的关联情况(见图 3)。

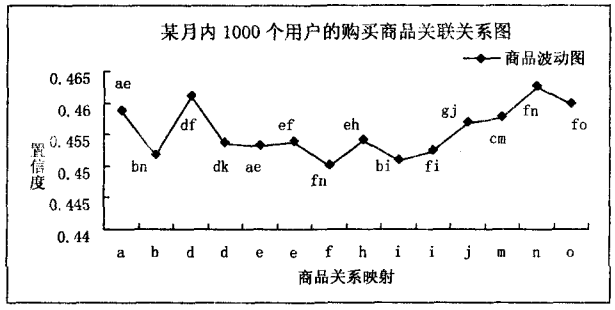

图 3　Apriori 算法得出某月内 1000 个用户的购买商品关联关系图

本次测试最小支持度设为 0.1,最小置信阈值设为 0.4,得出表 1 中的信息,以部分商品之间的关联关系说明一个月内销售的商品之间的关联关系,得出表 2 中的信息。

从两张表的比较中可知,顾客买了商品 n 再去买商品 f 的可能性很大;买了 o 的商品去买商品 f 的可能性也很大。买商品 n 和商品 o 的顾客都可能去买商品 f。

表 1　Apriori 算法得出 1000 个用户购买商品部分关系

商品 1	商品 2	置信度
n	f	0.462608
o	f	0.459864

表 2　置信阈值设为 0.4 时的部分商品关系信息

商品 1	商品 2	置信度
o	n	0.433333
f	n	0.450096
f	o	0.435286

综上所述,在 n,f,o 三中商品中顾客买了其中一种再买另外两种或两种之一的可能性比较大,超市管理者可以通过以上的结果可以调整商品摆放的位置,让这三种商品放在一起,方便顾客选购。有时候超市进行促销活动,则可能降低其中一种商品的价格,顾客买了促销的商品,很有可能就连带一起买其他两种商品,这样虽然降低了一种商品的价格,但是增加了其他商品的销售,也是超市盈利的一种很好的方法。

4　结束语

文中介绍了数据挖掘的相关概念,集中对 Apriori 算法进行研究应用。基于 Apriori 算法的理论知识,文

(下转第 107 页)

另外,实验证明基于信誉度的网格资源调度算法中,网格资源提供者为了获得更多的效益,将质量比较高的资源加入到网格中,相应的资源的信誉度就会随之提高,用户在调用资源时,直接根据信誉度的值选择比较符合自己要求的资源,从而大大节省了调用时间,同时还提高了任务执行的效率。

3　结束语

将信誉度引入到网格中,增强了用户在调用资源时的透明度,促使网格资源提供者将较高质量的资源引入到网格中来获取更高的效益,这样就达到优化网格质量的目的。另外,随着网格资源质量的提高,网格资源拥有者不断地将最优质量的资源加入到网格中来,扩大了网格规模。虽然用户调用资源的范围扩大,但时间缩短,减少了网格资源调度的时间,提高任务执行的效率。

但是,基于信誉度的网格资源调度算法还有一定的缺陷。假设,资源提供者提供的某些资源质量非常高,规定用户调用的时间及费用都比较合理,这样就会吸引大量的用户去调用这些资源,就很可能会造成严重的瓶颈问题。针对这一问题将做进一步研究。

参考文献:

[1] 马满福,吴　健,陈定剑,等. 网格经济模型中基于信誉度的资源选择[J]. 计算机工程,2006,32(17):175-177.

[2] 路　峰,吴慧中. 一种基于信誉 QoS 网格资源调度算法[J]. 信息与控制. 2009,38(2):170-175.

[3] Zacharia G, Maes P. Trust Management Through Reputation Mechanisms[J]. Applied Artificial Intelligence Journal,2000,14(9):881-908.

[4] 杨　柯,张建军. 基于计算期望和信誉度的网格资源调度模型[J]. 西北大学学报(自然科学版),2009,39(2):225-229.

[5] 李慧敏,蒋秀凤. 基于 QoS 效益函数的网格任务调度算法[J]. 计算机与现代化,2009(9):12-18.

[6] 王　进,解　福. 拍卖机制下的效益最优化调度算法研究[J]. 微计算机信息,2010,26 (4):205-207.

[7] 王　立,吴　蒙,常　莉. 移动 Ad hoc 网络基于信誉系统的节点协作方案[J]. 计算机技术与发展,2010,20(3):32-35.

[8] 穆晓芳,余雪丽,牛瑞萍. 基于拍卖机制的网格在线信誉系统模型[J]. 计算机工程与设计,2008,29(4):979-982.

[9] Buyya R, Murshed M. GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing[J]. Journal of Concurrency and Computation: Practice and Experience, 2002, 14 (13): 1175-1220.

[10] 高　强,刘　波. 关于网格模拟器的研究[J]. 计算机技术与发展,2010,20(1):100-103.

[11] Sulistio A, Yeo C S, Buyya R. Visual Modeler for Grid Modelling and Simulation (GridSim) Toollit[R]. Australia: Grid Computing and Distributed Systems (GRIDS) Lab, Dept. of Computer Science and Software Engineering, The University of Melbourne, 2003:1123-1132.

(上接第 103 页)

中设计并实现了一个原型系统,对超市商品的购买相关性进行了数据挖掘,从而实现对关联规则算法的研究与应用。

参考文献:

[1] Han Jiawei, Kammber 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, 2004.

[2] 彭小宁. 数据仓库与数据挖掘技术[J]. 怀化师专学报, 2002,21(2):34-38.

[3] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco: Morgan Kaufman Publisher, 2001.

[4] Agrawal R, Imiclinski T, Swami A. Mining Association Rules between Sets of Items in Large Database[C]//Proceedings of the 1993 ACM SIGMOD Conference on Management of Data Table of Contents. New York: ACM,1993:207-216.

[5] 周　涛,陆惠玲. 关联规则挖掘算法研究[J]. 齐齐哈尔大学学报, 2004,20(3):58-61.

[6] 毕建欣,张岐山. 关联规则挖掘算法综述[J]. 中国工程科学, 2005,7(4):88-94.

[7] Aflori C, Craus M. Grid implementation of the Apriori algorithm [J]. Advances in Engineering Software,2007,38(5): 295-300.

[8] Wang Yanhua, Feng Xia. The optimization of Apriori algorithm based on directed network[C]//Proceedings of the 3rd international conference on intelligent information technology application. Washington DC: IEEE Computer Socety,2009:504-507.

[9] Cheung D W L, Han Jiawei, Ng V. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique[C]//Proceedings of the Twelfth International Conference on Data Engineering. Washington DC: IEEE Computer Society,1996:106-144.

[10] 何丽君,董　蕊,袁克杰. 常见关联规则算法分析与比较[J]. 大连民族学院学报, 2005,7(5): 39-42.

[11] 吴芬兰,胡朝举,高　推. 关联规则挖掘算法的改进[J]. 微机发展(现更名:计算机技术与发展), 2005,15(8):151-152.

[12] 徐章艳,张师超,区玉明. 挖掘关联规则中的一种优化的 Apriori 算法[J]. 计算机工程, 2003,29(19): 83-87.