第 26 卷第 2 期　　四川理工学院学报(自然科学版)　　Vol. 26　No. 2
2013 年 4 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Apr. 2013
文章编号:1673-1549(2013)02-0036-03　　DOI:10.3969/j.issn.1673-1549.2013.02.009

基于关联规则分析的产品销售推荐的应用

李洪燕, 万　新

(四川理工学院自动化与电子信息工程学院, 四川　自贡　643000)

摘　要:阐述了关联规则算法的基本原理,并利用事务数据库中的销售数据和 SQL Server 2008 Data Mining Add – Ins for Microsoft office 2007 工具挖掘顾客购买的商品之间各种有趣联系,帮助商家制定营销策略,最终向每个顾客提供一组正确的推荐信息,从而改善客户的购物体验,增加总的销售额。

关键词: Microsoft 关联规则算法;销售推荐;客户细分

中图分类号:TU375　　**文献标志码:**A

交叉销售是实际生活中常见的商业问题,它是根据客户的购物篮中的历史产品信息来推荐客户最有可能购买的产品[1]。最优的推荐信息会提高顾客的购买欲,从而增加总的销售额。相反,劣质的推荐信息可能使客户失去购买兴趣。当面对的销售产品目录比较小时,根据丰富的销售经验来提供建议时可能比较容易实现,当不同产品的信息和数量比较庞大时,推荐信息就会失去有效性和准确性。

1　关联规则算法的基本原理

Microsoft 关联规则算法属于 Apriori 关联规则算法系列[2]。Microsoft 关联规则算法由两部分构成,第一部分是挖掘频繁项集,第二部分是基于频繁项集来生成关联规则。

1.1　关联规则算法的概念

设 $L=\{L_1, L_2, \cdots, Lm\}$ 是项的集合,设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是一个非空项集,使得 $T \subseteq L$。每一个事务都有一个标识符,称为 TID。设 A 是一个项集,事务 T 包含 A,当且仅当 $A \subseteq T$。关联规则是形如:$A \Rightarrow B$ 的蕴涵式,其中 $A \subset L$, $B \subset L$, $A \neq \Phi$, $B \neq \Phi$,并且 $A \cap B = \Phi$。

1.2　挖掘频繁项集

Microsoft 关联规则算法使用一种逐层搜索的迭代方法[3]。首先,通过扫描数据库,累计每个项的计数,并收集满足最小支撑度的项,找出频繁 1 项集的集合 L_1。然后,使用 L_1 找出频繁 2 项集的集合 L_2,使用 L_2 找出 L_3,依次下去,直到所有的侯选项集都不满足条件,算法终止[4]。

挖掘为事务数据库 D(图 1)的频繁项集的过程如下所示:

TID	商品ID的列表	TID	商品ID的列表
T_1	I1, I2, I5	T_6	I3, I4
T_2	I2, I3, I4	T_7	I2
T_3	I2, I4	T_8	I1, I2, I3
T_4	I1, I3, I5	T_9	I1, I2, I3, I4, I5
T_5	I1, I5	T_{10}	I3, I5

图 1　事务数据 D

(1)首先设每个项都是候选 1 项集的集合 C1 的成员。扫描所有的事务,并对每个项的出现次数计数。

(2)假设最小支持度计数为 2,即 minimum_support =2。确定频繁 1 项集的集合 L_1(图 2)。

(3)若想找到频繁 2 项集的集合 C2,首先连接 L1 产生候选 2 项集的集合。

(4)扫描 D 中事务,同时累计 C_2 中每个候选项集的支持度计数(图 3)。

(5)确定频繁 2 项集的集合 L2,L2 是 C2 中满足最

收稿日期:2013-01-08

作者简介:李洪燕(1986-),女,四川资阳人,硕士生,主要从事数据挖掘方面的研究,(E-mail)lihongyan623@tom.com

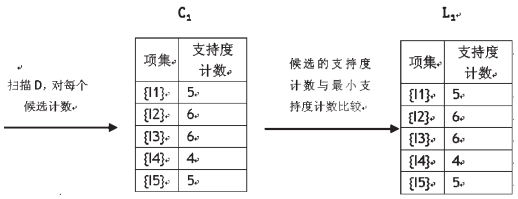

图 2　候选项集 C_2 和频繁项 L_2 的 产生过程

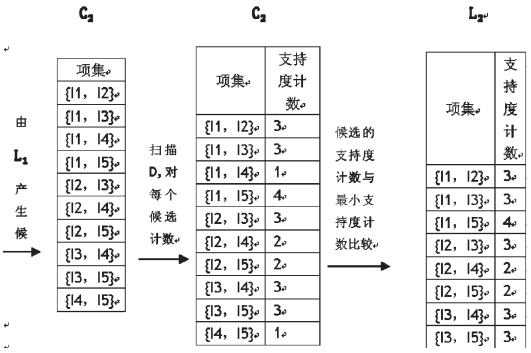

图 3　候选项集 C_2 和频繁项集 L_2 的生产过程

小支持度的候选 2 项集构成的。

(6)确定候选 3 项集的集合 C_3。首先令 L_2 = {{I1,I2,I3},{I1,I2,I5},{I1,I3,I5},{I2,I3,I4},{I2,I3,I5},{I2,I4,I5},{I3,I4,I5}},根据先验性质,频繁项集的所有子集必须是频繁的,可以确定{I2,I4,I5},{I3,I4,I5}不是频繁项集,因此,把它们删除。

(7)扫描 D 中事务以确定 L_3,它由 C_3 中满足最小支持度的候选 3 项集组成,如图 4 所示。

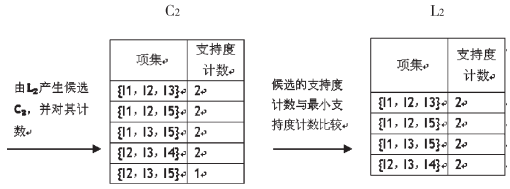

图 4　候选项集 C_2 和频繁项集 L_2 的生产过程

(8)候选 4 项集的集合 L_3 = {{I1,I2,I3,I5},{I2,I3,I4,I5}},但它的子集{I2,I3,I5}不是频繁的,所以 L_4 = Φ,算法终止,找出了所有的频繁项集。

2　事务数据源描述

在客户购买的历史信息中,提取了 32 265 个样本数,部分数据源如图 5 所示。图 5 中包含了客户的订单号、购买的产品的类别和单价。

3　数据分析

一旦提取完所需数据,并通过预处理,就可根据任务选择适当算法对数据进行分析处理。在该任务中,利用 SQL Server 2008 Data Mining Add－Ins for Microsoft office 2007 工具完成商品推荐,并选择关联算法对数据源进行分析处理,处理结果如图 6 所示。

Order Number	Category	Product	Product Price
SO61269	Helmets	Sport-100	53.99
SO61269	Jerseys	Long-Sleeve Logo Jersey	49.99
SO61270	Fenders	Fender Set - Mountain	21.98
SO61271	Tires and Tubes	LL Road Tire	21.49
SO61271	Tires and Tubes	Patch kit	564.99
SO61272	Tires and Tubes	Mountain Tire Tube	4.99
SO61272	Tires and Tubes	Patch kit	564.99
SO61273	Bottles and Cages	Water Bottle	4.99
SO61274	Caps	Cycling Cap	8.99
SO61274	Shorts	Women's Mountain Shorts	69.99
SO61275	Helmets	Sport-100	53.99
SO61276	Jerseys	Short-Sleeve Classic Jers	539.99
SO61276	Caps	Cycling Cap	8.99
SO61277	Mountain Bikes	Mountain-500	539.99
SO61277	Jerseys	Short-Sleeve Classic Jers	539.99
SO61277	Caps	Cycling Cap	8.99
SO61278	Road Bikes	Road-350-W	2443.35
SO61278	Bottles and Cages	Road Bottle Cage	8.99
SO61278	Bottles and Cages	Water Bottle	4.99
SO61278	Jerseys	Short-Sleeve Classic Jers	539.99
SO61279	Mountain Bikes	Mountain-200	2319.99
SO61279	Fenders	Fender Set - Mountain	21.98
SO61280	Helmets	Sport-100	53.99

图 5　事务数据

捆绑商品	捆绑大小	销售数量	销售平均值	捆绑销售总值
Road Bikes, Helmets	2	805	1570.228025	1264033.56
Mountain Bikes, Tires and Tubes	2	569	2208.067434	1256390.37
Fenders, Mountain Bikes	2	539	2022.477421	1090115.33
Mountain Bikes, Bottles and Cages	2	563	1923.73222	1083061.24
Mountain Bikes, Helmets	2	537	1966.57311	1056049.76
Jerseys, Road Bikes	2	480	2183.375083	1048020.04
Touring Bikes, Helmets	2	536	1925.792761	1032224.92
Road Bikes, Tires and Tubes	2	486	1541.535514	749186.26
Road Bikes, Bottles and Cages	2	552	1157.4025	638886.18
Touring Bikes, Bottles and Cages	2	351	1819.513846	638649.36
Jerseys, Mountain Bikes	2	284	1769.032817	502408.32
Touring Bikes, Tires and Tubes	2	205	1964.783707	402780.66
Touring Bikes, Jerseys	2	182	2140.195275	389515.54
Caps, Touring Bikes	2	192	1845.028125	354245.4
Mountain Bikes, Bottles and Cages, Helmets	3	172	2014.376977	346472.84
Caps, Road Bikes	2	218	1498.217339	326521.38
Caps, Mountain Bikes	2	193	1613.720933	311448.14
Gloves, Road Bikes	2	188	1579.83234	297008.48
Hydration Packs, Mountain Bikes	2	124	2197.621129	272505.02
Fenders, Mountain Bikes, Helmets	3	131	2015.929466	264086.76
Mountain Bikes, Helmets, Tires and Tubes	3	122	2010.395656	245268.27
Fenders, Jerseys, Mountain Bikes	3	102	2310.072745	235627.42
Fenders, Mountain Bikes, Bottles and Cages	3	114	2002.902982	228334

图 6　基于关联算法的详细分析信息

从图 6 中可知各类捆绑商品的销售数量、销售价格以及销售总值。以捆绑 road bike 和 helmets 为例:road bike 和 helmets 同时购买的销售数量为 805,从对事务数据的统计中可知客户单独购买 helmets 单价为 53.99,销售数量为 3794;购买 road bike 的单价为 2443.35,销售数量为 2369;平均销售单价为 1248.67 低于捆绑销售单价的平均销售价格 1570.22;销售总值为 1 005 179 也低于捆绑销售总值 1 264 033,因此利用该解决方案可增加产品的总销售额。

4　商品推荐

通过对数据进行分析和处理之后,就可得到理想的解决方案,从图 7 中推荐商品依据可知,如果客户购买了 fenders 就可向客户推荐 Mountain Bikes,并且购买该商品的购买率为 43.54%;客户购买了 Cleaners、Helmets、Bike Stands、Bike Racks 等任何一样产品就可向客户推荐 Tires and tubes,并且其中最有可能购买的客户为已购买了 Bike Stands 产品的客户;客户购买了 Gloves 可推荐 Helmets,购买率为 41.46%;客户购买了 Hydration

Pack 就可推荐 Bottles and cages 其购买率为 44.63%。因此如果市场部要策划一次营销活动,就可根据此解决方案来制定营销策略,有针对性地寄发产品海报,从而节约营销成本,得到最大的客户响应度和产品购买率。

所选商品	推荐	所选商品的销售情况	关联销售	关联销售的百分比	推荐的平均值	关联销售总值
Fenders	Mountain Bikes	1238	539	43.54%	870.97586	1078268
Cleaners	Tires and Tubes	525	259	49.33%	182.31973	95717.86
Helmets	Tires and Tubes	3794	1617	42.62%	9.4182789	35732.95
Bike Stands	Tires and Tubes	130	103	79.23%	243.19715	31615.63
Bike Racks	Tires and Tubes	191	94	49.21%	158.78325	30327.6
Gloves	Helmets	849	352	41.46%	22.384547	19004.48
Hydration Pack	Bottles and Cag	428	191	44.63%	3.8600234	1652.09

图 7 推荐商品的详细依据

5 结束语

庞大的销售数据库中隐含了客户的消费习惯和行为特征,而关联规则算法[5]能够帮助数据分析者和营销决策者发现海量交易数据背后的有价值的信息,以 SQL Server 2008 Data Mining Add – Ins for Microsoft office 2007 工具提供的样例数据基于 microsoft 关联规则算法建立模型来实现购物篮分析,并生成推荐信息,帮助商家制定营销策略,合理安排,提高销售额[6]。

参 考 文 献:

[1] 孙晓佳,朱宏丽.浅谈如何成功实践交叉销售[J].现代商业,2008(21):112.

[2] Crivat J M,著.董 艳,程文俊,译.数据挖掘原理与应用[M].北京:清华大学出版社,2010.

[3] Pei J H,著.范 明,孟小峰,译.数据挖掘概念与技术[M].北京:机械工业出版社,2012.

[4] Agrawal R, Shafer J C. Parallel mining of association rules: Design, implementation, and experience[M]. IEEE Trans. Knowledge and Data engineering, 1996.

[5] Ballou D P, Tayi G K. Enhancing Data Mining: Models and Algorithms[M]. New York: Springer, 2008.

[6] 徐 菊.商业性文献数据库的营销策略研究[D].广东:广东师范大学,2008.

Product Selling and Recommendation Application Based on the Analysis of the Association Rules

LI Hong-yan, *WAN Xin*

(School of Automation and Electronic Information, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: The basic principle of Microsoft Association rules algorithm is illustrated. Based on it, various fascinating relations among purchased goods are tapped by taking advantage of marketing data in the affairs database and SQL Server 2008 Data Mining Add-Ins for Microsoft office 2007 tools, which can help traders formulate marketing strategies and provide rational recommendation information for every customer. Thorefore, it can improve the shopping experiences and increase total sales.

Key words: Microsoft Association rules algorithm; selling recommendation; customer subdivision