关联规则算法在邮政商函客户关系中的应用

张志锋,邓璐娟,刘秀梅
(郑州轻工业学院 计算机与通讯工程学院,河南 郑州 450002)

摘 要:数据挖掘是当前数据库和信息决策领域的最前沿研究方向之一,在信息化技术发展的今天其重要作用十分明显。基于全新的信息技术的管理理念——客户关系管理受到中国邮政的青睐。数据挖掘技术在邮政商函CRM系统中起着核心作用,关联规则算法是数据挖掘的核心技术,在数据挖掘中是关键应用技术。文中在对关联规则算法和邮政商函客户关系分析研究的基础上,通过把关联规则算法运用在实例中,证明了关联规则算法在邮政商函客户关系管理起到一定的作用,有很好的应用前景。

关键词:数据挖掘;关联规则;客户关系管理
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)12-0238-03

Application of Association Rules Algorithm in Postal Customer Relationship

ZHANG Zhi-feng, DENG Lu-juan, LIU Xiu-mei
(Department of Computer and Communication Engineering, Zhengzhou Institute of Light Industry, Zhengzhou 450002, China)

Abstract: Data mining is the leading research area in the database and information decision. Data mining technology has core function in the postal CRM. The association rules algorithm is the core technology in data mining. On the basis of the analysis of association rule algorithm and the postal customer relationship management, and through the real application, it has been approved the association rule will play the significant role in the postal customer relationship management, and will have the good application in future.

Key words: data mining; association rules algorithms; customer relationship management

0 引 言

随意中国的改革不断深入,中国邮政的产业结构与市场环境发生了根本性的变化。中国邮政服务市场和业务逐步形成了从最初垄断市场到数家大跨国公司、许多家小公司不断加入的新格局。在一个全新的、更加竞争激烈的市场环境下,如何提升自身的核心竞争力,是一个关键问题。中国邮政人意识到,客户才是企业生存和发展的根基,而保存客户、吸引客户和充分发掘客户的盈收潜力是企业提高核心竞争力的关键。如何通过提高客户的满意度及忠诚度,提升客户价值来扩大自身的收入及利润等问题,成为中国邮政人关注的焦点。

在这种背景下,一种基于全新的信息技术的管理理念——客户关系管理(CRM, Customer Relationship Management)受到中国邮政的青睐。数据挖掘技术在邮政商函CRM系统中起着核心作用。目前,邮政系统已建立了邮政绿卡收费系统、邮政前台服务系统、邮政物流系统,通过这些系统可以方便地获得大量的客户数据。再通过建立邮政商函CRM系统就可以把所有与客户有关的数据进行整合成面向主题的数据仓库。然后,应用数据挖掘工具对这些数据进行挖掘以获得经营管理决策中所需要的信息和模式。这些信息和模式可以为企业的经营决策提供有力的决策依据。

文中结合数据挖掘和CRM两大技术,通过实例把它们结合起来,应用到实践中,可以提高邮政商函客户关系管理的效率。

收稿日期:2008-03-11
基金项目:河南省自然科学基金项目(0411010500),河南省新世纪优秀人才支持项目(2005HANCET-03)
作者简介:张志锋(1978-),男,河南郸城人,硕士,助教,主要研究方向为软件工程、数据库。

1 邮政商函客户关系管理

1.1 客户关系管理的基本概念

CRM是一个获取、保持和增加可获利客户的过程。CRM源于"以客户为中心"的新型商业模式,是一

种旨在改善企业与客户之间关系的新型管理机制,它通过将人力资源、业务流程与专业技术进行有效的整合,向企业的销售、市场和服务等部门和人员提供全面、人性化的客户资料,并强化跟踪服务和信息分析能力,使得企业可以以更低成本、更高效率满足客户的需求,并与客户建立起基于学习型关系基础上的一对一营销模式,从而可让企业最大程度地提高客户满意度及忠诚度,挽回失去的客户,保留现有的客户,不断发展新客户,发掘并牢牢地把握住能给企业带来最大价值的客户群。CRM 将先进的思想与最佳的实践具体化,通过使用当前多种先进的技术手段帮助企业从根本上提升核心竞争力,使企业在当前激烈的竞争环境中立于不败之地。

1.2　邮政商函客户关系管理的特点

邮政业务中主体业务是信函。其中以商业信函的数量占的份额较大,而且具有独特的市场特性和客户特性。首先,邮政企业的客户具有多元性,从党政机关、企事业单位、各种组织及社会团体,直到居民个人都是其客户;其次,邮政客户需求具有多元性,从团体至个人,从城市到农村,从低收入家庭到高收入家庭有各种层次的需求;最后,市场竞争性较强,客户使用邮政服务的随机性也较强。这决定了邮政企业的 CRM 有自己的特点和需求,80% 的利润来自占客户总量 20% 的企业客户;客户加入时间越长,对邮政的价值越高;老客户介绍新客户是最有效、最经济的销售方式;了解客户对邮政业务服务的需求才能推出满足客户需求的客户打包服务,提高客户的忠诚度并留住客户;目标客户的类别划分越明确,促销效果越好,转换率越高。

针对邮政业务的特点和竞争需求,要求邮政商函 CRM 分析的主要内容有:一是有关客户的年龄、收入、地区、性别、婚姻、种族、职业、职称和文化水平等;二是分析占比例最大和最小的客户群;三是评定客户信用度、排名贡献;四是分析客户风险系数及对经营风险的影响程度;五是分析客户的流失情况等。

2　关联规则算法

关联规则是数据挖掘的核心技术,它是由 Agrawal 等人首先提出的。关联规则就是给定一组项目(Item)和一个记录集合,通过分析记录集合,推导出 Item 间的相关性。关联规则广泛地应用于商业界、医疗保险、金融业、司法部门等,因此对它的研究有着极其重要的意义。

在关联规则系统中,规则本身是"如果条件怎么怎么样,那么结果或情况就怎么样"的简单形式。关联规则可表示为 A⇒B。左部 A 称为前件,右部 B 称为后件。前件可以包括一个或多个条件,在某个给定的正确率中,要使后件为真,前件中的所有条件必须同时为真。后件一般只包含一种情况,而不是多种情况[1,2]。

以正确率(也称置信度)为目标的关联规则,主要是以正确率表示前件为真时后件为真的可能性。对于"A→B"关联规则,其置信度可定义为:

置信度(A→B)=包含 A 和 B 的元组数/包含 A 的元组数

对用户来说最重要的是规则的正确率。正确率达到 80% 以上的规则,表明发现的关系是很强的。即使它们对数据库的覆盖率较低,出现的次数不多。

以覆盖率(也称支持度)为目标的关联规则表示数据库中适用于规则的记录数量。可定义为:覆盖率(A→B)=包含 A 和 B 的元组数/元组总数。覆盖率高表示规则经常被使用。

关联分析可数学形式化地描述为:

定义 1 设 $I=\{i_1,i_2,\cdots,i_m\}$ 是由 m 个不同的属性(谓词或项目)组成的集合(习惯上还称 I 为项集,但其中的元素与 Agrawal 等人的定义有所不同,这里项集中的元素可能是谓词或项目,而 Agrawal 等人定义的项集仅包含项目)。给定一个数据库 D,其中的每一个记录 T 是 I 中一组属性的集合,即 T 包含于 I。若集合 X 包含于 I 且 X 包含于 T,则记录 T 包含集合 X。一条关联规则就是形如 $X\rightarrow Y$ 的蕴涵式,其中 X 包含于 I,Y 包含于 I,$X\cap Y=\varnothing$。关联规则 $X\rightarrow Y$ 成立的条件是:

(1) 它具有支持度 S。即在数据库 D 中至少有 $S\%$ 的记录包含 $X\cup Y$;

(2) 它具有置信度 C。即在数据库 D 中包含的 X 记录至少有 $C\%$ 的同时也包含 Y。习惯上将关联规则表示为 $X\rightarrow Y(S\%,C\%)$。

支持度:$S\%$ = The Number of Transactions($X\cup Y$)/The Number of Transactions(D)

置信度:$C\%$ = The Number of Transactions($X\cup Y$)/The Number of Transactions(X)

其中,支持度定义了项目在整个数据库中所占的比例;置信度定义了发现规则的强度[2,3]。

3　应用实例

下面结合客户寄发商业信函实例提出一个可行的关联分析方法在客户关系管理中的实际应用。

某些房地产公司对武汉市城区的用户进行寄发商业信函来宣传他们的房屋,邮政函件广告公司对一定时间范围内客户寄发区域详细情况作了收集,情况如

表1所示(限于篇幅,仅列出6个客户、5个地域)。

表1 客户发函情况表

客户	发函区域
A	江汉区、江岸区、桥口区
B	武昌区、桥口区、汉阳区
C	江岸区、桥口区
D	武昌区、汉阳区、江岸区、江汉区
E	武昌区
F	武昌区、汉阳区

针对表1进行关联分析,首先构造两种地区间的关联表,如表2所示,表中每一个数值表示的是行、列代表的两种区域同时被一个房地产商寄发的次数。

表2 两个地区间的联系表

Y \ X	江岸区	江汉区	桥口区	武昌区	汉阳区
江岸区	3	2	2	1	1
江汉区	2	2	1	1	1
桥口区	2	1	3	1	1
武昌区	1	1	1	4	3
汉阳区	1	1	1	3	3

针对设定的最小支持度阈值,计算每一个 x 的最小支持度(本例,设最小支持度阈值为0.3):support(江岸区,桥口区)=0.33;support(江岸区,江汉区)=0.33;support(武昌区,汉阳区)=0.5。其他未列出。

针对设定的最小置信度阈值,计算最小置信度,如表3所示。

表3 $X-Y$ 最小置信度

Y \ X	江岸区	江汉区	桥口区	武昌区	汉阳区
江岸区	/	0.667	0.667	0.333	0.333
江汉区	1.0	/	0.5	0.5	0.5
桥口区	0.667	0.333	/	0.333	0.333
武昌区	0.25	0.25	0.25	/	0.75
汉阳区	0.333	0.333	0.333	1.0	/

将大于最小置信度阈值的列出(本例,设最小置信度阈值为0.7),即为关联分析所得出的规则:

Rule1:江岸区→江汉区,support=0.33,confidence=0.667

Rule2:江岸区→桥口区,support=0.33,confidence=0.667

Rule3:桥口区→江岸区,support=0.33,confidence=1.0

Rule4:江汉区→江岸区,support=0.33,confidence=0.667

Rule5:武昌区→汉阳区,support=0.5,confidence=0.75

Rule6:汉阳区→武昌区,support=0.5,confidence=1

从上述规则可以初步得出结论:

(1)房地产公司客户中相当比例的在向江汉区发函的同时几乎肯定要向江岸区寄发。说明大部分的房地产商在江汉区、江岸区这两地区发函有群带效果。

(2)房地产公司客户中相当比例的在向武昌区发函的同时几乎肯定要向汉阳区寄发。

根据上述规则,公司在营销时采取了如下措施:

1)在向房地产商介绍时将江汉区与江岸区、武昌区与汉阳区信息数据在一起推荐。

2)营销员在客户选择一个地区后,适当推荐另一个地区。

(3)客户保持。市场竞争越来越激烈,使企业获得新客户的成本正不断上升,因此保持原有客户就显得越来越重要。客户分为3类:第1类是无价值的客户;第2类是不会轻易走掉的有价值的客户;第3类是为不断地寻找更优惠的价格和更好的服务的有价值的客户。传统的市场活动是针对前两类客户的,而现代客户关系管理认为,特别需要用市场手段来维护的客户是第3类客户,这样做会降低企业运营成本。数据挖掘可以发现易流失的客户,企业就可以针对客户的需求,采取相应措施。

(4)一对一营销。CRM系统可以把大量的客户分成不同的类,在每个类里客户拥有相似的属性,而不同类里的客户属性也不同。企业可以做到给不同类客户提供完全不同的服务来提高客户的满意度。细致而切实可行的客户分类对企业的经营策略有很大益处,数据挖掘可以帮助企业针对不同类别的客户,提供个性化的服务。

4 结束语

用数据挖掘中的关联规则算法可以很好地发掘邮政商函市场的潜力,降低市场运营成本,稳定客户资源,从而提高市场的竞争力和服务质量。如果使关联规则算法能够更好地发挥其作用也要依赖于商业数据的采集和准确性,以及完整的CRM系统的建立特别是数据仓库系统和流程的建立,对于应用各种工具产生的结论还要得到业务专家的确认或评估。相信正确运用数据挖掘技术会使邮政商函CRM发挥更大的作用,真正成为提高邮政主体函件业务的有力武器[4~6]。

参考文献:

[1] 陆建江.加权关联规则挖掘算法的研究[J].计算机研究与发展,2002(10):1281-1286.

[2] 齐 雁.对演变数据进行关联规则挖掘的新方法[J].计算

(下转第244页)

$$p_{ij}=\begin{cases}1 & GS_i \quad \text{path } GS_j \\ 0 & GS_i \quad \text{not path } GS_j\end{cases} \tag{2}$$

上式中 GS_i path GS_j 表示 GSRG 中 GS_i 到 GS_j 有一条路。

定理 5 PGSR 中若按服务关系的顺序依次对服务编号，即最小服务的编号为 1，最大服务的编号为 n，PGSR 是 TGSR 当且仅当 PGSR 的可达性矩阵为一对角线元素全为零而其它元素都为 1 的上三角矩阵。

证明 TGSR 中，对于任何 $GS_i(i=1,2,\cdots,n-1)$，均能为所有 $GS_j(i<j)$ 提供服务，而所有的 $GS_k(i>k)$ 均不能为其提供服务，根据 GSRG 的可达性矩阵的定义，可得以上结论。

定理 6 PGSR 中，GS_i 是最小服务当且仅当 G 的可达性矩阵中 $P_{ji}=0 \wedge P_{ij}=1(j=1,2,\cdots,n \wedge i\neq j)$，$GS_i$ 是最大服务当且仅当 PGSR 的可达性矩阵中 $P_{ij}=0 \wedge P_{ji}=1\ (j=1,2,\cdots,n \wedge i\neq j)$。

证明 根据最小服务的定义，最小服务能为其它任何服务提供服务，因此 i 行除 $P_{ii}=0$ 外其余元素都为 1，而其它任何服务不能为最小服务提供服务，因此第 i 列全为 0。同理，根据最大服务的定义，其它任何服务都能为最大服务提供服务而最大服务不能为其它任何服务提供服务，可得以上结论。

上面的两个定理给出了通过服务关系的可达性矩阵判定全序网格服务关系和偏序网格服务关系中的最小服务和最大服务的方法。在网格环境中，对于给定的一个输入的作业，不能再分解的的任务集合通常是偏序集，在某些情况下可能是全序集，通过任务集到服务集的映射，由任务集的偏序关系可动态驱动服务偏序集的执行。如果最先执行的任务是一个，通过映射得到的服务集合中必然存在最小服务，就可以自动确定最先执行的服务，如果最后执行的任务是一个，也必存在最后执行的最大服务。在最小任务和最小任务中间执行的任务，可以是顺序的或并行的，可以由任务偏序集驱动服务的动态组合。

3 结束语

文中的贡献在于首次从二元关系的角度对网格服务关系及其性质进行了研究，为任务集与服务集之间的映射、服务之间的匹配、服务组合和服务协同等的研究提供了理论的支持。在文中研究的基础上，下一步的工作重点是研究基于二元关系的服务自动匹配和服务动态组合。

参考文献：

[1] Foster I, Kesslman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations[J]. The International Journal of High Performance Computing Application, 2001, 15(3): 200 - 222.

[2] Foster I, Kesslman C, Nick J M. The physiology of the grid - An open grid services architecture for distributed systems integration[EB/OL]. Open Grid Service Infrastructure WG, Global Grid Forum. 2002. http://www.globus.org/research/papers/ogsa.pdf.

[3] GGF, Open Grid Services Infrastructure (OGSI) version 1.0 [DB/OL]. 2003. www, 2003.

[4] Czajkowski K, Ferguson D F, Foster I, et al. The WS - Resource Framework, Version 1.0[EB/OL]. 2004 - 03. http://www - 106. ibm. com/developerworks/library/ws - resource/ws - wsrf. pdf.

[5] 易 明，金 海. 基于 WSRF 的 Web 服务资源的设计[J]. 计算机工程，2006，32(23)：262 - 263.

[6] Li Maozhen, Baker M. 网格计算核心技术[M]. 王相林，张善卿，王景丽，译. 北京：清华大学出版社，2006：21 - 25.

[7] 左孝凌，李为镒，刘永才. 离散数学[M]. 上海：上海科学技术文献出版社，1982：139 - 145.

(上接第 237 页)

社，2003.

[2] 张桂元，贾燕枫. Eclipse 开发入门与项目实践[M]. 北京：人民邮电出版社，2006.

[3] 王国辉，王 易. JSP 数据库系统开发案例精选[M]. 北京：人民邮电出版社，2006.

[4] 徐国智. SQL Server 数据库开发实例精粹[M]. 北京：电子工业出版社，2006.

[5] 赫斯特. 实战 STRUTS[M]. 北京：机械工业出版社，2005.

[6] 易枚根. Dreamweaver8 网页设计与网站建设[M]. 第 2 版. 北京：机械工业出版社，2007.

(上接第 240 页)

机工程，2002(11)：126 - 127.

[3] 尹阿东. 基于数值属性的关联规则挖掘算法[J]. 微机发展，2003，13(4)：67 - 70.

[4] 周 欣. 兴趣度——关联规则的又一个阈值[J]. 计算机研究与发展，2000(5)：627 - 633.

[5] 王兴鹏. 面向 CRM 的数据挖掘应用[J]. 计算机时代，2003(3)：44 - 46.

[6] Gibson J P. Formal requirements models: simulation, validation and verification[R]. Maynooth: National University of Ireland, 2002：132 - 140.