第35卷 第20期 计 算 机 工 程 2009年10月
Vol.35 No.20 Computer Engineering October 2009
·软件技术与数据库· 文章编号：1000—3428(2009)20—0081—02 文献标识码：A 中图分类号：TP311

基于数据挖掘的供热用户分析与决策

卢铮松[1]，赵 洁[2]
(1. 天津大学管理学院，天津 300072；2. 天津市房产供热公司，天津 300192)

摘 要：对某供热公司累积的大量供热用户收费数据进行分析，通过构建数据仓库和利用数据概化方法建立供热用户数据挖掘模型，使用频繁项集方法产生关联规则，利用决策树算法得出交费时间特征，从而得出不同区域和类型用户的习惯交费时间段。对该数据挖掘模型进行评价，提出的4项收费决策建议在实际应用中取得良好效果。

关键词：集中供热；数据挖掘；用户管理

Analysis and Strategy for Customers of Heat Supply Based on Data Mining

LU Zheng-song[1], ZHAO Jie[2]
(1. Management School, Tianjin University, Tianjin 300072; 2. Tianjin Real Estate Supply Heating Company, Tianjin 300192)

【Abstract】 This paper analyzes a great deal of charge data that a supply heating company provides. It establishes a data mining model for customers by constructing a data warehouse and using the data generalization, and finds out some association rules from the frequent itemsets and some features of time to pay by using decision tree, so that it can draw a conclusion that different customers have different custom of charge. It evaluates the model and gives four suggestions on the charge strategy used in actuality.

【Key words】 central heat supply; data mining; customers management

1 概述

城市集中供热广泛推行之后，作为经济实体的供热公司实现了自主经营，自负盈亏，因此，供热收费成为各供热公司一项非常重要的工作。许多供热公司都投入了大量资金，购买或开发了完整的用户交费系统。在经过几年的运行后，这些用户交易系统中积累了大量的交费信息数据和用户数据。为了能充分利用这些积累的记录信息，从中发现有用的知识，从而提高信息利用率，更合理地调配有限的资源，更有效地进行供热收费，需要实现经营分析系统。而实现分析系统的主要技术就是数据仓库和数据挖掘。

数据挖掘(data mining)是指从大量数据中提取或发现知识[1]。数据挖掘通过一些模型和智能方法，从大量数据中提取数据模式，并根据某些兴趣度度量，识别出用户真正感兴趣的、新颖的、潜在有用的模式，提供给用户作为决策的依据和参考。作为关键技术之一，数据挖掘已经在很多方面得到应用，尤其是在经营分析系统中，对于客户潜力资源的开发、市场的细化分析、战略决策调整等方面，都发挥了重要的作用。

利用数据挖掘从大量数据中发现知识的过程有如下步骤：(1)定义分析主题，进行数据预处理；(2)设计数据模型，进行数据变换和选择；(3)综合使用多种数据挖掘方法，建立数据挖掘模型；(4)对挖掘出的模式进行评估和实施[2]。

本文对供热用户交费信息分析下的交费习惯与用户住址之间的联系进行了描述。通过对一个典型的供热公司自2000年~2006年的用户交费数据进行数据挖掘，得到用户习惯交费时间段与住户片区之间的逻辑关系。运用这个规律进行用户交费预测，供热公司可合理地安排收费人员的分配、制定收费策略、调整优惠政策等。

2 主题定义分析

根据任务，需要通过数据挖掘研究的是住在不同区域用户的交费习惯。因此，需要执行的数据挖掘是一个关联规则。对于与此主题相关的数据及背景知识分析如下。

2.1 关系数据库的基本属性

在供热公司的交费管理系统中，通常采用关系数据库系统作为后台数据库。在本文所研究的供热公司的收费系统中，关系数据库包括十几个数据表和上百个字段。不失一般性，在通常的供热交费管理系统中，存在表1的属性。其中，交费年度为用户所交本笔费用应为某一年度的供热费用，而交费时间为用户交费的具体日期。由于供热费按照单位面积收取，因此表中较为明显的关系是C3->P1。

表1 关系数据库

所属关系表	规模	字段
用户信息表	中等	用户名 C1 用户住址 C2 住房面积 C3
用户交费表	较大	交费金额 P1 交费年度 P2 交费时间 P3

2.2 用户交费优惠政策

为了鼓励用户主动交费，根据政府主管部门的政策，在供热收费工作中，通常有一些相关的优惠措施出台。目前的优惠政策Y的规定是：每年6月1日~6月30日交当年供热费，可优惠5%，7月交费可优惠4%，依次递减，直至10月交费优惠1%，11月之后则全额交费。因此，交费金额和交

作者简介：卢铮松(1977－)，男，助理研究员、博士研究生，主研方向：人工智能，数据挖掘；赵 洁，工程师

收稿日期：2009-03-20 **E-mail：**Lzspyc@tju.edu.cn

费时间有对应关系 P3->P1。

2.3 数据仓库的构建

数据仓库和 OLAP 工具基于多维数据模型。该模型将数据看作数据立方体(data cube)形式。数据立方体由维和事实定义。根据相关性分析，给出与客户交费记录相关的维度表与其事实表的星型模型，如图 1 所示。

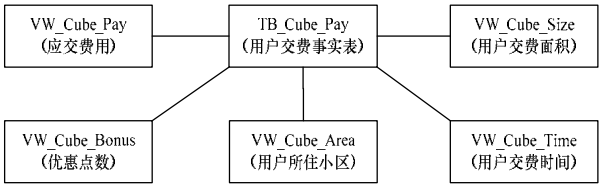

图 1 星型模型

3 数据概化和模型建立

数据概化是把数据从低层次抽象到高层次的过程，包括 2 种方法：数据立方体方法(第 2 部分)和面向属性的归纳。

3.1 基于交费时段的泛化

由于交费月份和交费用户数之间的关系是可以量化的，为了描述基于不同月份的交费情况，可采用二维表或条形图的形式来表示，见表 2、表 3。

表 2 2005 年不同月份交费情况的概化关系

月份	交费用户数	占交费人数的比例
6	6 148	0.321 716
7	1 264	0.066 143
8	222	0.011 617
9	347	0.018 158
10	792	0.041 444
11	2 252	0.117 844
12 月及之后	5 864	0.306 855
本年度未交费	2 221	0.116 222

表 3 2006 年不同月份交费情况的概化关系

月份	交费用户数	占交费人数的比例
6	7 943	0.415 646
7	1 116	0.058 399
8	195	0.010 204
9	327	0.017 111
10	986	0.051 596
11	2 025	0.105 965
12 月及之后	6 178	0.323 286
本年度未交费	340	0.017 792

3.2 基于汇总的用户分布特征化

对于用户地址的分布情况，通过对用户地址进行聚类分析，可将地址信息分解为小区、楼栋、楼层等信息，按照数据立方体的划分方法进行划分。

同时，住宅面积(即收费面积)与用户地址分布也有密切的关系。一些小区或一些楼栋户型较小，另一些小区或楼栋则户型较大。根据地址信息和住宅面积的条件进行聚类分析，可以从数据中得到一些任务相关的集合。

根据所在城市住宅特点，可针对住宅面积进行概化，将住宅面积根据如下规则划分为 5 个等级：

(1)30 m^2 以下：一室小户型；

(2)30 m^2~70 m^2：一室一厅小户型；

(3)70 m^2~110 m^2：二室一厅中户型；

(4)110 m^2~150 m^2：三室一厅大户型；

(5)150 m^2 以上：别墅或超大户型。

对以上分类进行属性相关性分析，因为小区、楼栋和住户地址之间有强相关关系，所以在此后的考察中，需要选取其中一个作为建立模型的数据。为简化起见，本文选择小区层次进行关联挖掘。需要对楼栋进行关联挖掘时，可执行数据挖掘的下钻操作。

4 关联规则挖掘

对原始数据进行整理和特征化变换后，即可对其进行数据挖掘。本文使用改进的多维关联算法，设定客户的交费时间为规则目标。在找到的频繁项集中，先考虑可信度。设定支持度为 1%，可信度为 40%，按照如下步骤进行关联挖掘：

(1)根据交费数据，预选属性的集合。

在所考察的数据中，选择 2005 年和 2006 年的数据作为训练集，预选相关的属性。所选取的属性包括用户所属小区、交费面积、交费时间、交费金额、优惠金额等。

(2)产生频繁项集。

取训练集中的记录和记录的属性值形成矢量化矩阵。矩阵中列为属性，行为属性值，构造所有属性为元素的候选集 C1(即 1 个属性的组合)。计算 C1 的支持度，去掉 C1 中支持度小于 1%的属性，得到频集 L1。根据 L1 及 2 个属性的组合，构造候选集 C2，计算 C2 的支持度，去掉 C2 中支持度小于 1%的二维属性，得到频集 L2。依次类推重复上述过程，直到所有属性组合完毕，形成频繁项集 L={L1,L2,…}。

(3)由频繁项集产生关联规则。

根据频繁项集生成可信度大于 40%的规则，若规则太多且难以归类，则应调整支持度和可信度，重新进行挖掘，剔除掉明显不合理的规则，最终得到强关联规则。其中一部分强关联规则见表 4。

表 4 住户与交费倾向的部分强关联规则

强关联规则	支持度/(%)	可信度/(%)
用户面积小=>12 月交费	2.4	65.1
用户面积大=>6 月交费	3.1	61.2
小区规模大=>12 月交费	1.4	55.7
高层=>12 月交费	2.1	60.3

(4)使用决策树算法，找出用户交费时间的特征。

从以上数据集中，根据不同供热片区客户的特点，选用扩展的 C5.0 算法作为基本算法[3]，采用自顶向下的递归算法，构造出决策树模型。

算法 DTree(dmset, attributes)中的 DTree 为递归构造函数；dmset 为数据集；attributes 为记录集。使用信息增益比度量法的启发式规则定义熵，从而描述信息增益比。熵越小，则属性值的分布差别越大；熵为 0 时，则成为叶节点。在实际运用中，每次取熵最小的属性作为递归构造节点。

在构造过程中，需要不断对决策树进行修剪。在构造某节点时，若该节点下的记录树在所有记录树中的比例小于 0.1%，则停止该子树的生长。决策树生成后，若可信度小于 30%，则剪掉该树枝。决策树生成后，可通过验证直到误差满足要求为止。从决策树上可以看出不同小区、不同户型、不同年龄层次之间用户交费的特征。

5 结束语

本文通过对提供居民集中供热公司收费数据的分析，构建了一个数据挖掘的模型。该数据模型挖掘了收费优惠时段和用户之间存在的一些强关联关系。通过数据挖掘模型，可以得到收费策略的决策树，并依此制订相应的收费策略。经过检验和总结，得出的如下一些收费策略：

(1)对于以小户型为主的旧小区，住户多为中低收入家庭，应积极宣传优惠条件，并加强上门收费工作，减少不交费情况的发生。

(下转第 85 页)