第18卷 第5期
2008年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 18 No. 5
May 2008

支持CRM分析的数据仓库多维启动模型

闫娜娜,刘 锋,李锡娟,耿 波
(安徽大学 计算机学院,安徽 合肥 230039)

摘 要:CRM是一种整合了知识管理、数据挖掘及数据仓库技术的商业策略,旨在支持制定决策来保留长期有利的客户关系。分析了CRM中数据仓库的设计问题,提出一种支持CRM分析的强劲的多维启动模型。为验证此模型,用一些CRM的查询进行测试,并定义两个量:成功率和适配率来评估。实验结果表明,此启动模型具有很高的成功率和适配率,可用于客户收益分析、市场收益分析、产品收益分析、渠道收益分析等多种收益分析。

关键词:数据仓库;CRM;多维启动
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)05-0067-03

A Multidimentional Starter Model of Data Warehouse to Support CRM Analysis

YAN Na-na, LIU Feng, LI Xi-juan, GENG Bo
(Department of Computer Science and Engineering, University of Anhui, Hefei 230039, China)

Abstract: CRM is a strategy that integrates knowledge management, data mining, and data warehousing in order to support decision-making process to retain long-term and profitable relationships with customers. In this paper, analyze the data warehouse design to support CRM and propose a multidimentional starter model that supports CRM analysis. Present sample CRM queries, test the model using those queries and define two measures - success ratio and suitability ratio to evaluate the model. The experiment result shows that success ratio and suitability ratio of the model are high, and the model can be used to analyze customer profitability, market profitability, product profitability, channel profitability and so on.

Key words: data warehouse; CRM; multidimentional starter

0 引 言

目前CRM及其相关技术受到越来越多的关注。但支持CRM的数据仓库设计还没有统一的标准。CRM数据仓库模型的设计直接影响着企业执行分析的能力,关系着CRM的成败[1]。为了客观地评估提出的CRM模型,定义两个量:成功率(success ratio)和适配率(suitability ratio)。

成功率取值范围是(0,1),值越大说明模型越成功。

$$r_{success} = q_p / q_n \tag{1}$$

其中q_p是在此模型上被成功执行的查询数,q_n是查询总数。

适配率取值范围是(0,1),取值越大,说明模型适配性越好。

$$r_{suitability} = \sum_{i=1}^{n} x_i c_i / n \tag{2}$$

其中n代表可应用分析标准的总数,c代表每个分析性能的得分,x代表每个分析性能被赋予的权重。用这两个量来评估提出的CRM数据仓库模型。

1 CRM数据仓库模式设计

CRM数据仓库模式设计的第一步先识别出CRM相关分析的不同类别。特殊的兴趣数据点可以凭经验识别出来[2]。表1中列出需要考虑的相关分析的类别及数据维护问题。也就是CRM数据仓库的最低设计要求。

1.1 事实表设计的基本原理

此模型由收益事实表、未来价值事实表、客户服务事实表及表2中定义的各个维度表组成。收益事实表包含的属性是计算每笔交易的历史收益所必需的最小集。这样可以提高数据仓库的查询能力。另外,逐条存储交易可以提高计算在每种产品中每个客户的客户

收稿日期:2007-08-27
基金项目:安徽省自然基金项目(070412051);安徽高校省级重点自然科学研究项目(KJ2007A43)
作者简介:闫娜娜(1983-),女,硕士研究生,研究方向为数据挖掘;刘 锋,教授,硕士生导师,研究方向为并行分布计算。

终身价值(CLV)的能力[3]。另外,此模型还可用于计算投递 KPIs,例如准时投递项的数目及无损坏项的数目。补足量可由总量减去明确存储的 KPI 量得到。KPIs 在追踪和管理方面很重要,因为可以帮助企业识别出可以改进处理的区域,从而影响客户满意度及客户保持。

表 1 CRM 数据仓库的最低设计要求

分析类型/数据维护	描 述
Customer Profitability(客户收益)	确定各个客户收益的能力
Product Profitability(产品收益)	确定各产品收益的能力
Market Profitability(市场收益)	确定各市场收益的能力
Campaign Analysis(活动分析)	评估不同的活动及随时间响应的能力
Channel Analysis(渠道分析)	评估各渠道收益的能力
Customer Retention(客户保持)	追踪客户保持的能力
Customer Attrition(客户流失)	识别客户流失原因的能力
Customer Scoring(客户评分)	给客户评分的能力
Household Analysis(家庭分析)	将客户和扩展家庭账目相联系的能力
Customer Segmentation(客户细分)	将客户细分到相应客户集的能力
Customer Loyalty(客户忠实度)	理解不同关系组中忠实模式的能力
Demographic Analysis(人口统计分析)	执行人口统计分析的能力
Trend Analysis(趋势分析)	执行趋势分析的能力
Product Delivery(产品投递)	评估准时/提前/推迟产品投递的能力
Product Returns(产品返还)	分析产品返还原因的能力
Customer Service Analysis(客户服务分析)	追踪和分析客户满意度/互动成本/解决客户投诉时间等的能力
Up-selling Analysis(提升销售分析)	分析客户购买大量或高利润产品可能性的能力
Cross-selling Analysis(交叉销售分析)	识别客户可能购买的附加产品的能力
Web Analysis(网络分析)	分析网站客户访问量的能力
Data Maintenance(数据维护)	维护客户细分及评分的历史记录的能力
Data Maintenance(数据维护)	从不同数据源整合数据的能力
Data Maintenance(数据维护)	有效地更新/维护数据的能力

客户服务事实表包括每一次与客户互动的信息,比如,互动成本、解决投诉的时间、客户满意率或不满意率等。客户总的历史价值可以由每笔交易产生的价值(存储在收益事实表中)总和减去总的客户交互成本(存储在客户服务事实表中)。

1.2 维表设计的基本原理

维度是根据事实表中要用以分析的维确定的。每个维在确定前要根据下列原则检验:

(1)同一维中如果存在有一些属性跟其余的属性变化率不同;

(2)包含有整个历史价值都需要维护的属性;

(3)从属于间断存在(比如只在特殊时期应用)。

如果上述情况任何一种成立,那么就提取出来单独建维。此外,情况(1)中,新的维在 outriggers 中可作为 mini 维实现,使得用户可快速浏览事实表。另外此方法的优点是客户行为评分和人口统计变化的历史记录会作为事实表的一部分存储,这样可以使分析更加灵活、强劲。

因为事实表只能捕获交易发生时的历史价值,但这是一个静止期,在此期间发生的任何变化都不会记录到数据仓库中去。这将会影响那些本可以做的分析,因为一旦没被记录,就不能分析数据。所以在设计步骤要给予充分考虑。

建立客户维和销售代表维、市场维、注释维及时间维之间的直接对应关系可使用户通过简单的浏览客户维,就能容易地确定销售代表、市场、活动日期、流失日期及流失注释项的取值,而不需要在查询中包含时间约束。

表 2 启动模型的维度定义

维度名称	维定义
Channel Dimension(渠道维)	存储与客户互动的不同方式
Customer Dimension(客户维)	存储客户的静态信息
Customer Behavior Dimension(客户行为维)	存储客户的动态评分属性
Customer Demographics(客户人口统计)	存储客户动态的人口统计特征
CustomerExistence(客户存在)	追踪客户有效期
CustomerMarket(客户市场)	追踪客户和市场维之间关系的变化
Comments Dimension(注释维)	存储客户流失和产品返还的原因
Company Representative(企业代表)	存储企业代表及销售代表
Extended Household(家庭扩展)	表示客户可能属于一个或多个扩展家庭关系
Market Dimension(市场维)	客户从属的组织层次或区域
Product Dimension(产品维)	表示企业所售产品
ProductExistence(产品存在)	追踪产品的有效期
Promotion Dimension(促销维)	表示企业的促销活动
Scenario Dimension(细节维)	用以分析假设的提升销售和交叉销售
Supplier Dimension(供应商维)	表示产品的卖主
sTime Dimension(sTime 维)	模式中统一的时间标准
Time Dimension(时间维)	模式中统一的日期标准
County Demographics Dimension(地区人口统计维)	存储地区的人口统计

2 实 验

文中用实验来验证启动模型的有效性。此 CRM 数据仓库模型运行平台是 SQL Server 2000[4]。此模型的收益事实表由 1 685 809 条数据记录组成。

2.1 方 法

在实验中,基于提出的数据仓库模式执行一系列的 CRM 查询。成功率由成功执行的查询数除以实验中用到的查询总数。此外,此模型通过测试,确定表 1 中的各个任务能否执行。如果能执行,就给 1 分,否则给 0 分。这些分数相加之和就得到 CRM 分析能力总分数。

2.2 用以测试的查询的选择

文中提出的数据仓库模式会对执行 CRM 分析的能力有积极的影响。因此,为使测试模型中用到的查

询具有普遍的代表性,实验中用到的查询用层次随机取样法选择[5]。层次随机取样方法如下:

(1)CRM中用到的有代表性的查询是从不受约束的集合中获取的;

(2)根据查询的自身性质将其分成不同的类;

(3)在每一个类中,每一个查询都是有编号的;

(4)随机数产生器用来从每个类中选择查询;

(5)如果查询对应的编号与随机数产生器产生的数字一样,那么这个查询就被选中了。被选中的查询及其得分如表3所示。

表3 CRM分析实例

种类	分析	pass	Fail
Channel Analysis	Which distribution channels contribute the greatest revenue and gross margin?	1	0
Order Delivery Performance	How do early, on time and late order shipment rates for this year compare to last year?	1	0
Order Delivery Performance & Channel Analysis	How do order shipment rates (early, on time, late) for this year compare to last year by channel?	1	0
Customer Profitability Analysis	Which customers are most profitable based upon gross margin and revenue?	1	0
Customer Profitability Analysis	What are the customers' sales and margin trends?	1	0
Customer Retention	How many unique customers are purchasing this year compared to last year?	1	0
Market Profitability Analysis	Which markets are most profitable overall?	1	0
Market Profitability Analysi:	Which products in which markets are most profitable?	1	0
Product Profitability Analysis	Which products are the most profitable?	1	0
Product Profitability Analysis	What is the lifetime value of each product?	1	0
Returns Analysis	What are the top 10 reasons that customers return products?	1	0
Returns Analysis	What is the impact of the value of the returned products on revenues?	1	0
Returns Analysis	What is the trend for product returns by customers by product by reason?	1	0
Customer Attrition	What are the top 10 reasons for customer attrition?	1	0
Customer Attrition	What is the impact of the value of the customers that have left on revenues?	1	0

要注意的是,因为这些查询是从与CRM相关的查询集中随机选择的,成功率可能小于提出的模型的成功率。另外,CRM代表性的查询是指在不同行业都能应用的查询而不是只能在一个行业中使用的特定查询。取样程序在概括不同行业执行CRM分析的数据仓库模式特性的方面非常重要。

2.3 实验结果

文中提出的数据仓库模型可成功地执行CRM分析,且基于样本查询得出模型的成功率和适配率分别为1和0.93。表3中列出了各个查询在模型上成功执行的得分。由于篇幅关系,各查询的适配率得分不一一列举。

2.4 实验查询分析

此启动模型只需要在查询语句块中包含一些合适的维度,即可用于多种CRM分析。另外,每一个查询都可以通过包含附加量及事实表和维度中的附加域的描述来提升查询能力。客户收益SQL语句可以识别客户总历史价值。它可以和客户未来价值,客户服务互动成本一起用来进行客户细分,从而用来确定相应的客户管理策略。

客户收益分析查询如下所示:

Which customers are most profitable based upon gross margin and revenue?

```
SELECT b. CustomerKey, b. CustomerName, Sum(a. GrossRevenue)
AS TotalRevenue, Sum(a. GrossProfit) AS Total GrossProfit, Total GrossProfit/
TotalRevenue AS GrossMargin
FROM tblProfitabilityfactTable a, tblCustomer b
WHERE b. CustomerKey = a. CustomerKey GROUP BY b. CustomerKey, b.
CustomerName
ORDER BY Sum(a. GrossRevenue) DESC
```

产品收益分析SQL语句可用于确定产品的利润,从而也可用于识别潜在的可能停产的产品。产品收益分析查询如下所示:

Which products in which markets are most profitable?

```
SELECT
c. Year, b. marketkey, b. LocationCode, b. Location, b. Description, b. Com petitorName, d. ProductCode, d. Name,
Sum(a. GrossRevenue) AS TotalRevenue, Sum(a. GrossProfit) AS Total GrossProfit,
    Total GrossProfit/ TotalRevenue AS GrossMargin
FROM tblProfitabilityfactTable a, tblMarket b, tblTimeDimension c, tblProduct-Dimension d
WHERE b. MarketKey = a. MarketKey and a. TimeKey = c. TimeKey and a. ProductKey = d. ProductKey
GROUP BY c. Year, b. CompetitorName, d. ProductKey, d. ProductCode, d. Name, b. MarketKey
ORDER BY Sum(a. GrossRevenue) DESC
```

仅通过修改上述SQL语句,除去产品代码就可以确定每一个产品的终生价值的能力,进一步说明了该模型的灵活性和强进性。

3 结 语

分析了CRM中的数据仓库设计问题,提出一种支持CRM的分析的多维启动模型。根据抽样查询实验得到的模型的成功率和适配率分别为1和0.93。此模型还可通过在SQL语句中包括或排除时间维来灵活地用于分析趋势、客户终身价值、市场、客户及产品。另外此模型可以捕获丰富的描述性的非数值信息,可以包含在查询语句中,所以模型返回的结果很容易被用户理解。并且这种丰富的信息可作为类标签在数据挖掘算法中使用。

(下转第72页)

表 1 某城市住房调查表

日期	住宅区域	房型	收入	前结果	现结果
3/10/06	近郊	分离	高	无应答	无应答
14/9/06	近郊	分离	高	已应答	无应答
2/4/05	农村	分离	高	无应答	已应答
18/1/06	城市	半分离	高	无应答	已应答
3/4/06	城市	半分离	低	无应答	已应答
15/10/05	城市	半分离	低	已应答	无应答
15/10/05	农村	半分离	低	已应答	已应答
2/3/04	近郊	综合	高	无应答	无应答
4/5/06	近郊	半分离	低	无应答	已应答
2/1/06	城市	综合	低	无应答	已应答
3/10/06	近郊	综合	低	已应答	已应答
3/10/06	农村	综合	高	已应答	已应答
8/4/06	农村	分离	低	无应答	已应答
6/5/05	城市	综合	高	已应答	无应答

表 2 收入和决策属性关系表

	已应答	无应答	总计
高	3	4	7
低	6	1	7
总计	9	5	14

接着根据表 2 计算收入属性的卡方值：

$$\chi^2_{income}=\frac{(3-3.5)^2}{3.5}+\frac{(4-3.5)^2}{3.5}+\frac{(6-3.5)^2}{3.5}+\frac{(1-3.5)^2}{3.5}=3.7$$

同理可以得到其他属性的卡方值，从表 3 中可以知道，日期属性的卡方值最大，所以就选择日期属性为决策树当前结点的分裂属性。

表 3 3 种方法分裂属性选择结果

	信息熵	增益比率	卡方值
日期	0.600	0.180	9.333
住宅区域	0.246	0.156	4.400
房型	0.049	0.031	2
收入	0.151	0.151	3.7
前结果	0.048	0.048	2

为了好比较，在表 3 中列出了实验过程中其他两种方法(即信息熵方法和增益比率方法)的结果。从以上三种方法得到的结果来看，虽然“日期”属性在三种方法中都被选为当前结点的分裂属性，但是明显的卡方检验方法区别能力更强(各种属性的卡方值分别是 9.333、4.4、2、3.7 和 2)，而信息比率区别能力最差(各种属性的卡方值分别是 0.180、0.156、0.031、0.151 和 0.048)。

4 实验分析

为了显示文中方法的优越性，从 UCI[5]数据库下载了 8 个数据库，把这 8 个数据库用上述三种方法在软件 C5(可以在 Quinlan 的主页上免费下载到)进行分类比较，然后比较这三种方法的分类错误率，实验结果显示在表 4 中。

表 4 8 个 UCI 数据集的实验结果

	Ecoli (%)	Breast (%)	Heart (%)	Thyroid (%)	Australia (%)	Mushroom (%)	Voting (%)	Cars (%)
信息熵	26.33	27.63	21.35	19.63	26.54	24.58	20.25	18.52
增益比率	25.31	28.53	19.56	22.81	25.84	25.34	21.63	16.25
卡方检验	20.23	22.89	18.23	17.65	25.39	24.59	19.30	15.62

5 结 语

基于统计学理论，提出了利用卡方检验的方法在决策树中进行分裂属性的选取。通过实验证明，文中的方法是非常有效的，并且卡方检验方法简单实用，应用起来非常方便，一些计算器和一些常用的软件例如 EXCEL 上都有计算卡方值的功能。

参考文献：

[1] Mitchell T M. 机器学习[M]. 曾华军，张银奎译. 北京：机械工业出版社，2003.

[2] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986(1):81－106.

[3] Quinlan J R. C4.5: program for machine learning[M]. New York, US: Morgan Kaufmann, 1993.

[4] Hunt E B, Marin J, Stone P T. Experiments in Induction [M]. New York, US: Academic Press, 1966.

[5] Blade C L, Merz C J. UCI repository of machine learning databases(website)[D]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

(上接第 69 页)

参考文献：

[1] Myron D, Ganeshram R. The Truth About CRM Success & Failure[J/OL]. CRM Magazine, 2002. http://www.destinationcrm.com/articles/default.asp? ArticleID=2370.

[2] Kellen, Vince. CRM Measurement Frameworks[EB/OL]. 2002. http://www.kellen.net/crmmeas.htm.

[3] Dipak J, Siddhartha S. Customer Lifetime Value Research In Marketing: A Review and Future Direction[J]. Journal of Interactive Marketing, 2002, 16(2):34－46.

[4] 沈兆阳. SQL Server 2000 OLAP 解决方案——数据仓库与 Analysis Services[M]. 北京：清华大学出版社，2001.

[5] 巴斯蒂安 M. 数据仓库与数据挖掘[M]. 北京：冶金工业出版社，2003:175－177.