数据挖掘项目的特征和关键环节

陈华英
(中国民用航空飞行学院,四川 广汉 618307)

摘 要:数据挖掘技术作为企业信息技术应用的自然延伸,正在成为近年来企业在实施数据仓库项目后的关注重点。文中在大量数据挖掘项目的实施总结基础上,对数据挖掘项目的特征、人员构成和角色分析、方法论和关键环节进行了深入分析。为以后不断地跟踪最新的数据挖掘知识和项目实施方法论,不断地通过数据挖掘项目实践来创造业务效益,提供了理论依据。数据挖掘相关理论和技术研究应该作为国内信息技术领域在今后一个时期的焦点命题。

关键词:数据挖掘;数据挖掘方法论;SEMMA; CRISP-DM; 记分
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2006)09-0085-03

Features and Key Processes of Data Mining Project

CHEN Hua-ying
(Civil Aviation Flight University of China, Guanghan 618307, China)

Abstract: Data mining technology attracts increasingly focus as the natural expansion of the IT application in today's enterprises. Deeply analyze the features, team member roles, methodology and key processes of data mining project, based on the experiences and summaries of projects implemented. It offers the academic gists for advanced track the knowledge of data mining and projects implement and create benefit. The correlative theory and technology of data mining must be focus proposition of domestic information technology domain.

Key words: data mining; data mining methodology; SEMMA; CRISP-DM; Scoring

1 数据挖掘简介

根据数据挖掘业界权威 Michael J. A. Berry 和 Gordon S. Linoff 的论述:数据挖掘是利用自动或半自动手段揭示大量数据中有意义的潜在规律的处理过程。这里需要强调的是"大量数据"和"有意义的潜在规律",这两个特征将数据挖掘与传统的独立分散的数据分析及简单的数据库查询、报表应用区分开来。

数据挖掘应用在近年来迅速发展,其基础是关系型数据库系统应用的逐步普及和成熟,以数据库形态存在的业务数据大量积累,为数据挖掘中的"大量数据"和"自动或半自动手段"提供了可能;其驱动力是业务需求的发展,尤其是数据库应用系统上线后给业务需求带来的正反馈作用;其核心是产品化的数据挖掘产品和实施咨询服务。

2 数据挖掘项目形态

2.1 基于数据仓库的数据挖掘

在很多项目中,数据挖掘是整合数据平台特别是数据仓库的延伸应用。通常,大型项目中,在数据仓库中为特定主题的数据挖掘建立数据集市,使得数据可以通过比较系统的形式定期加载更新,作为较为稳定的数据挖掘数据源;经过数据挖掘得到的数据规律,以计分预测或者与营销系统整合等形式发布到企业中,并经过一定的收效评估和阶段回顾,得出项目阶段性结论[1]。这种类型的项目,数据挖掘和数据仓库紧密结合,取用统一数据,有利于数据挖掘过程在企业的重用和固化,建立稳定的应用模式;但是数据挖掘的过程在较大程度上受到数据仓库建设的制约,见效的周期可能会较长,短期的投资见效比不理想,而且项目有很可能因数据仓库方面的问题而非数据挖掘的问题导致失败。

2.2 先导型数据挖掘

数据挖掘项目也可以独立于数据仓库存在。在挖掘的主题已经明确而相应的数据仓库还未建立,或者是项目有较强的预研性的情况下,数据挖掘项目可以直接进入主题,取用运营系统的原始数据,建立针对具体数据挖掘用途的专用数据区,不考虑太多的重用批量加载环节,尽快地开始挖掘过程,并将结果与业务迅速沟通。这样做的好处是便于企业更直接地体验数据挖掘的效益,尤其是业务管理部门可以很快得到来自数据规律的直接决策支持信息,数据挖掘受数据仓库建设过程的制约较少,见效周期短,短期的投资见效比比较好。但是比较难形成较为稳定的应用模式,同时由于数据源及转换处理往往独立于企业数据仓库建设,部分工作可能会在以后的数据集市过程中重复开始,甚至出现数据的不一致性,如果存在过多的这

收稿日期:2005-12-16
基金项目:中国民航飞行学院科研基金(J2004-23)
作者简介:陈华英(1968-),女,四川广汉人,副教授,硕士,研究方向为数据库。

种彼此独立的项目，将造成局部“信息孤岛”现象[2]。

在笔者参与实施的数据挖掘案例中，将以上两种模式有机地结合在一起，先利用一个或几个主题的独立数据挖掘项目的开展，为企业数据仓库提供面向数据挖掘的数据需求，同时，这些独立项目中的数据准备环节充分考虑数据仓库的思路。这样，在数据仓库建设中，可以得到更多的来自数据挖掘的设计要求和参考经验，有效地建立数据仓库和数据挖掘整体系统。

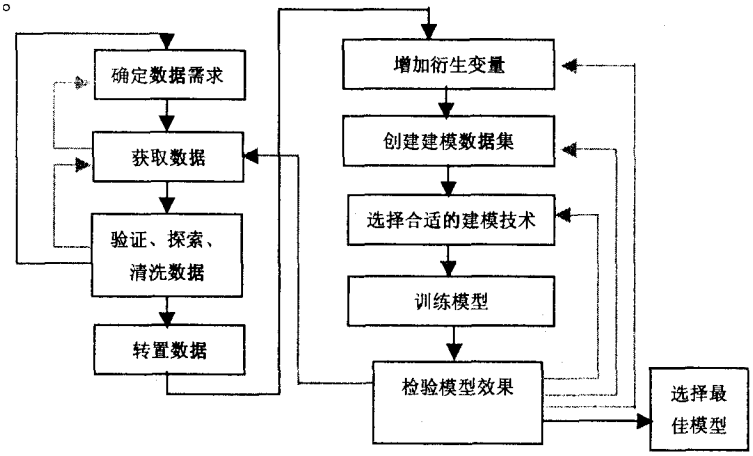

图 2 数据转换流程图

3 数据挖掘项目的架构

3.1 数据挖掘方法论简介

数据挖掘的架构是建立在成熟、合理的方法论基础上的。主要有 SEMMA 方法论和 CRISP－DM 方法论。SEMMA 方法论以抽样(Sample)、探索(Explore)、修改(Modify)、建模(Model)、评估(Assess)为核心环节，强调数据挖掘过程是这 5 个环节的有机循环。CRISP－DM 是跨行业数据挖掘标准流程(Cross－Industry Standard Process for Data Mining)的缩写，强调以业务理解(Business understanding)、数据理解(Data understanding)、数据准备(Data preparation)、建模(Modeling)、评价(Evaluation)、发布(Deployment)为核心环节，将数据挖掘目标和商务目标有机地联系在一起[3]。

在实际应用中，将上述两种方法有机地结合起来，CRISP－DM 强调高层的商务目的实现过程，SEMMA 强调具体的数据挖掘技术实现过程。

3.2 主要环节

综合实际进行的数据挖掘，数据挖掘项目可以分为以下几个主要环节，如图 1 所示。

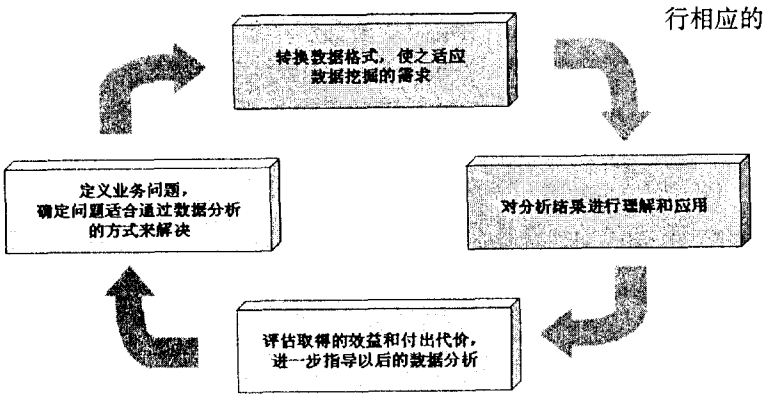

图 1 数据挖掘项目的主要环节

1) 定义业务问题。

这个环节的任务包括：评估数据挖掘过程的成本和商务收益间是否平衡；识别分析目标的焦点范围；收集相关的业务规则；确定数据源的可用性和验证行业专家的观点。

2) 转换数据格式使之适应数据挖掘的要求。

这是技术性最强的环节，包括了数据准备和数据挖掘建模。主要流程如图 2 所示。

a. 确定并获取数据。首先，要根据已经明确的业务问题，定义需要被预测或研究的目标因素。然后，确认数据中包含在历史上已经发生的目标因素的结果值，例如，预测客户流失，历史数据中需要包含客户是否发生流失的信息。同时数据中还应该包含与目标因素可能相关的各类信息，在了解数据源的过程中，还应该明确数据的更新加载方式，这样才能够形成不断使用最近数据，预测未来目标的循环应用模式。

b. 验证，探索，清洗数据。需要确定数据的来源是否可靠。考察数据项是自动衍生还是手工录入；是否存在缺失现象；取值是否符合规定，是否合理；数值分布是否可以解释，等等。

c. 转置数据，形成合适的颗粒度。数据挖掘需要的数据往往是一个事件一行，一行中包含所有的相关属性。如客户价值分析中，以客户号为核心，将客户的各种指标在时间上的快照聚集到一行上。这种形式，需对原始数据进行相应的转置操作，例如，将多个属性行对应一个客户的结构转置成一个客户行多个属性列的格式。

d. 增加衍生变量。很多情形下，原始的数据列和目标因素之间不易找到明显的相关性，需要增加一些衍生变量，以辅助分析。例如，在客户使用量这个指标的基础上，增加客户的用量的三个月平均变动率，等等。

e. 准备建模用的数据。这个环节需要考虑分析的时间段和时间颗粒度(周、月、季等)，建模用的数据必须匹配相应的时间要求，数据中时间的发生必须在相应的时间段内。同时，可能需要对小概率事件进行过抽样(Oversampling)以适应建模技术。在很多情形下，还可能对数据做剖分(Partition)，将历史数据分为训练(Train)、验证(Validate)、测试(Test)3 个部分，以便取得较好的预测效果，避免过拟合(Overfitting)现象[4]。这些操作，将使数据更加适合数据挖掘的建模工作。

f. 选择合适建模技术，训练模型。这个环节，就是狭

(下转第 90 页)

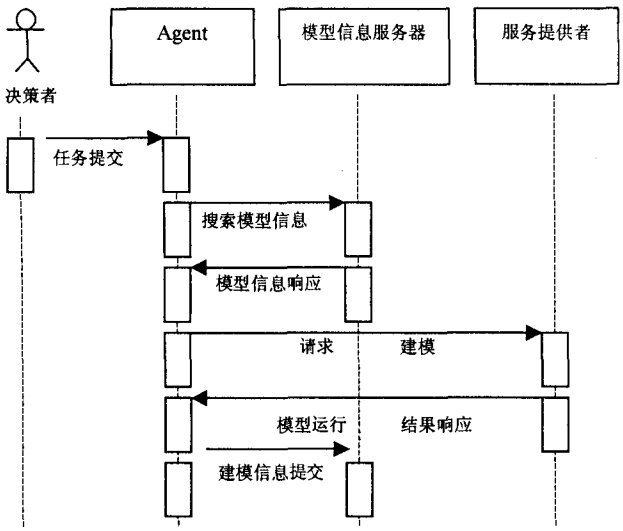

图 2 模型管理流程图

5 结束语

文中重点讲述了一种基于网格环境的模型状态管理的思路,并提出了一个模型服务技术体系框架。在网格新规范 WSRF 下,提出了用网格属性文档来管理模型的查询、建立、实施等模型生命周期的各个部分。

模型服务在分布环境下的性能和模型服务标准也是需要考虑的问题。性能一直以来就是网格计算的优势,所以随着网格技术和硬件条件的发展它会进一步的提高。模型服务则需要建立标准。而模型输入输出参数格式可以有很多种,各个机构都可以建立自己的模型。这些都会给模型管理带来困难。使用文中提出的模型属性文档来描述模型的思路是在建立模型标准的路上迈出了关键的一步。

参考文献:

[1] 洪一帆,荣 冈.面向对象的模型管理系统[J].科技通报,2002,18(6):446-450.

[2] 林 杰,雷星晖.基于 Web 服务的分布模型管理系统的研究[J].计算机应用,2004,24(4):80-82.

[3] Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure[M]. San Fransisco, CA: Morgan Kaufmann, 1999.

[4] Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations[J]. International Journal Supercomputer Applications, 2001, 5(3): 200-222.

[5] Czajkowski K, Ferguson D F, Foster I. WS-Resource Framework[EB/OL]. http://www.globus.org/wsrf/, 2004.

[6] Ong M, Ren X, Allan G, et al. Decision Support System on The Grid[A]. In proceedings of the Int'l Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES)[C]. New Zealand: Wellington Institute of Technology, 2004.

(上接第 86 页)

义上的"数据挖掘",实质上是挖掘建模的具体技术过程。采用 SEMMA 方法论逐步找到合适的建模技术,训练数据,最终找到规律和模式[5]。

g.检验模型的效果。在模型检验中,会使用历史数据中部分已有结果,以测试数据的形式与模型预测结果对比,客观地考察预测准确性。在真正的预测期间,只能等到未来的数据结果变成现实后,才能对预测结果作出对比,因此,需要有一个模型在市场环境中的试投放的时期,来检验模型真实效果。

3) 对分析结果进行理解和应用。

利用数据挖掘的最终结果和中间结果,可以深入了解企业数据的分布特征和存在的问题,进行一次性的专题分析或是周期性分析预测,还可以建立实时评分系统,如客户信用评分系统等,也可以为企业数据系统的改进提供重要的依据。

4) 评估模型的收效。

将模型的结果和投入成本与真实的业务收效相比,最终对数据挖掘过程作出综合评价。

4 小 结

数据挖掘项目在目前,特别是在国内,还处于边界条件尚未明确界分的阶段,并不是很成熟。但是数据挖掘项目的特质之一就是动态性,这种动态性是由它与企业业务的密切结合决定的,它对于业务的辅助作用的力度和直接程度超过了传统的业务支撑系统、MIS 系统,也超过了数据仓库应用中的报表查询系统;企业对于决策信息的需求,在数据挖掘项目中,找到了前所未有的载体,因此,数据挖掘应用拥有更加广阔深远的前景。随着数据挖掘中某些应用的进一步成熟,数据挖掘将在各大行业中逐步形成有层次的产业链。

所以,不断地跟踪最新的数据挖掘知识和项目实施方法论,不断地通过数据挖掘项目实践来创造业务效益,应该作为国内信息技术领域在今后一个时期的焦点命题。

参考文献:

[1] Berry M J A, Linoff G S. Mastering Data Mining[M]. [s. l.]: John Wiley & Sons, 2000.

[2] Berry M J A, Linoff G S. Data Mining Techniques[M]. [s. l.]: John Wiley & Sons, 1997.

[3] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.

[4] 萨师煊. 数据库系统概论[M]. 北京:高等教育出版社,2004.

[5] 郭崇慧. 数据挖掘教程[M]. 北京:清华大学出版社,2005.