# 数据挖掘商业应用平台的设计与实现

## Design and Implementation of Data Mining Platform for Commercial Use

刘绍清 （福州职业技术学院 350002）
黄章树 （福州大学管理学院 福建福州 350002）

**摘要**：本文设计并实现了建立于B/S与三层C/S混合架构上的，非专业用户能够轻松使用的通用的数据挖掘商业应用平台。在需求分析的基础上[1]，着重探讨了平台的逻辑结构、模块功能划分以及平台的实现与技术特点，给出了平台逻辑结构图和模块结构图。

**关键词**：数据挖掘 商业应用平台 设计与实现 逻辑结构

本文重点探讨一个不需要使用者面对复杂的算法，不需要掌握众多非常专业的数据挖掘工具及高超的技巧，同时又不昂贵的通用数据挖掘辅助工具：数据挖掘商业应用平台的设计与实现。

## 1 平台的需求分析

因为数据挖掘所面对的往往是海量数据，多样化的、分散的、内容经常不一致的数据，需要用复杂的算法，反复探索这些数据，从中找出一些不为人知的知识，这就要求参与者要非常精通各种算法，非常熟悉数据，对参与者提出了非常高的要求，所以，数据挖掘商业应用平台实现的功能必须能够帮助用户在不需了解算法细节的情况下自如地使用各种算法；能够保证数据定义一致、数据间关系清晰明确，用户能轻松、清楚、全面地了解和掌握企业数据资产的信息，了解原始数据如何经过一步一步的处理达到当前的状态；能够把知识以用户可以理解的方式展示，并把展示结果可以保存成网页的形式，以供在互联网上发布结果，或存为Excel文件形式以便进一步的加工处理。

为此，该平台应该具备以下功能：

元数据管理的功能，对元数据从生成到应用到退出的整个生命周期进行管理；用可视化的方式灵活地展现一个数据挖掘项目所涉及到的数据内容、数据含义以及数据之间的联系；从数据源抽取小部分数据样本的取样功能，并且对样本的操作能够大体展现操作的结果，但是由于数据规模小很多，平台响应将快的多，在最终确定数据操作之前的探索过程中等待的时间将大幅度减少，从而达到提高数据挖掘效率的目的；用可视化地方式对数据清洗和整理，从而大大地简化数据挖掘的数据准备过程；提供挖掘结果的展示功能；提供用户自定义数据挖掘算法的功能；通过定义一个流程实现整个数据挖掘流程的多个阶段，改变以往一个数据挖掘过程需要采用多种工具配合、一个过程被迫分割成若干个不连贯的处理流程的状况，从而降低数据挖掘过程管理难度和对使用者的要求。

## 2 平台逻辑结构设计

在前面需求分析的基础上，可设计出如下的平台逻辑结构：

整个平台由以下几部分构成：

第一部分是数据挖掘所涉及的业务数据来源，为数据挖掘提供原始的数据，这些业务数据可能是以关系数据库的形式、文件的形式、甚至纸质材料的形式存在。

第二部分是商业应用服务器和数据库服务器，它能够定义元数据抽取规则、ETL（数据抽取、转换、装载）规则存入元数据库，定义商业逻辑存入模型库中；平台将根据这些规则自动抽取元数据，执行ETL操作，业务数据经过抽取、清洗、转换、集成等预处理手段，消除了噪声，去除了不一致，最后被存放在数据挖掘库中，为后续的OLAP与数据挖掘提供高质量的数据资源，保证OLAP与数据挖掘不需要直接面对原始数据；平台能够根据用户指令对数据挖掘库中的数据进行分析，或者

根据商业逻辑进行 OLAP 和查询，并把结果传递给第三部分的客户端程序，展示给用户。

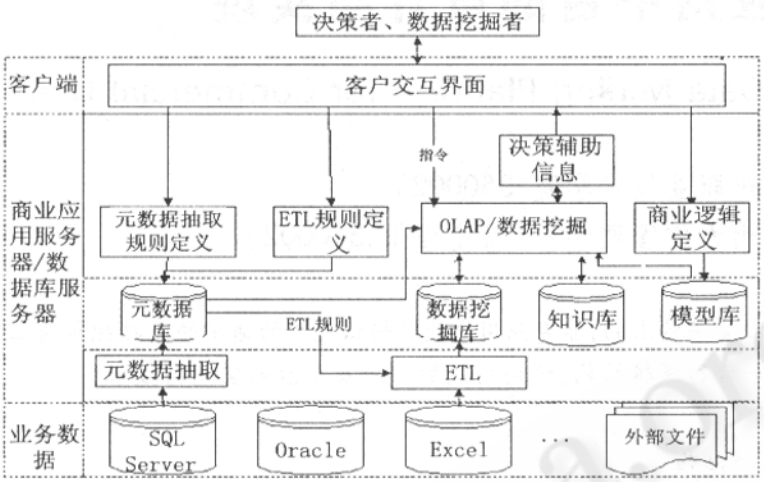


图 1　平台逻辑结构示意图

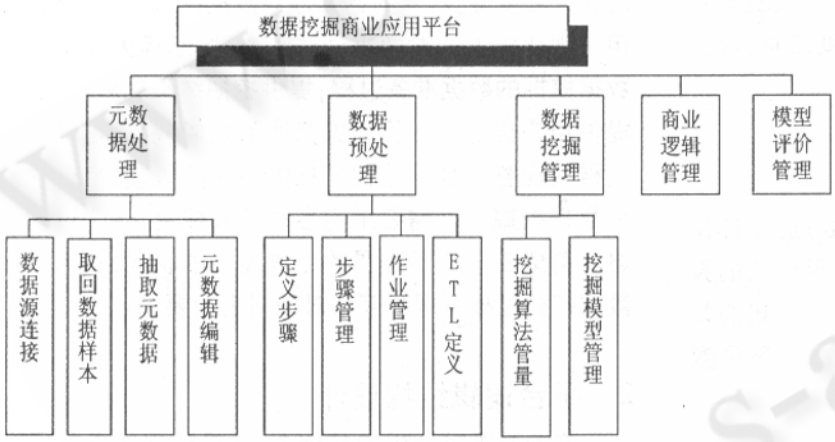


图 2　平台模块结构图

这部分实现的功最终能都被包装成功能单一、高聚合低耦合的应用服务器，以方便部署和提高性能。

第三部分是客户端软件，它以使用者熟悉的形式提供可交互的图形接口供用户定义各种规则和商业逻辑，定义数据挖掘的目的和意图，展示数据挖掘结果、OLAP 结果以及查询结果，以期达到降低对平台使用者要求的设计目标。

## 3　平台功能模块设计说明

整个平台按功能上分成元数据管理、数据预处理管理、数据挖掘管理、商业逻辑管理、模型应用管理 5 个子系统，每个子系统又分为若干模块，如图 2 所示。

各个模块的功能具体阐述如下：

### 3.1　元数据管理

元数据也就是“关于数据的数据[3]”，它和数据仓库（数据库）的关系就像图书目录和图书馆的关系[4]。元数据管理子系统在平台中处于中心位置，其他各个部分都是围绕着它实现的。由于数据挖掘项目的背景各异、挖掘目的不同，实施过程中要接触的数据源类型、规模、数量、使用的时机会随着项目的不同而不同，即使在同一项目中也会随着项目的进展发生变化。所以，本子系统要实现以下功能：

能够可视化地设置数据库连接信息，一旦环境发生变化的时候，用户能够通过修改设置适应这种变化；能够自动抽取源数据的元数据，并对元数据的使用、调度、引退等方面进行管理，为平台收集的元数据添加注释和显示信息，以方便和统一用户的理解，为后续的数据挖掘过程提供统一规范的信息，保证数据挖掘团队在整个数据挖掘过程中对同一内容有统一的理解；能够在抽取元数据的同时，抽取源数据的一个随机样本，在最终确定数据操作之前的探索过程以这个规模小的多的样本集为操作对象，对样本的操作能够大体展现操作的结果，但系统响应时间将比直接对数据源操作的响应时间短很多，在最终确定数据操作之前的探索过程中等待的时间将大幅度减少，从而达到提高数据挖掘效率的目的。具体分为数据源连接、取回样本数据、抽取元数据、元数据编辑四个模块。

### 3.2　数据预处理

数据预处理包括数据清洗、集成、转换和消减[5]。数据清洗是指消除数据中所存在的噪声以及纠正其不一致的错误。数据集成就是将来自多个数据源（如：数据库、文件等）的数据按照统一的格式结合在一起并形成比较完整的数据集合，为数据挖掘的顺利完成提供数据基础。数据转换主要是对数据进行规格化操作，将数据转换或归并以构成一个适合数据挖掘的描述形

式。数据消减是在不影响(或基本不影响)最终的挖掘结果的情况下,大幅度缩小所挖掘数据的规模,从而大幅度减少后面数据预处理和数据分析所耗费的时间。这四种数据预处理不是相互独立的,而是相互关联的,通过灵活组合这几种数据预处理方法,能够有效地帮助改善数据的质量,进而帮助提高数据挖掘进程的有效性和准确性[6]。

数据预处理子系统分为步骤定义、步骤管理、作业管理、ETL 定义四个子模块,提供了 10 类 15 个预处理节点,每类分别侧重某一方面的数据处理技术[2],而且还允许用户定义自己的节点,使用者可用箭头将这些节点串起来形成的链条表示一个数据预处理流程。

用户可以根据数据本身的特点和不同数据挖掘算法对数据的需求,选择合适的节点,有机地组合各种预处理方法,用可视化的方式进行清洗、集成、转换和消减等预处理工作内容,在提供高质量数据的同时大大地简化数据挖掘的数据准备过程。

另外,数据预处理过程中一些过程需要多次反复,修改是常事,如果在多个地方有重复定义的相同或相似的操作,一旦需要修改,则很可能花了很多时间逐个修改,却还是可能有几个地方被遗漏了没有修改,或者是几个不同的地方的修改不完全一样,这个问题不是靠花更多时间可以解决的。预处理子系统允许用户将相同或相似操作设置成一个相对独立的步骤存放在元数据库中,在 ETL 过程中重复使用,步骤定义一旦修改,则所有应用这个步骤的地方都会自动修改,极大地方便了数据预处理过程的维护,提高数据预处理过程管理效率。

### 3.3 数据挖掘模型生成管理

这个子系统分为 2 个部分:数据挖掘算法管理和挖掘模型管理。

鉴于数据挖掘算法非常多,涉及到机器学习、统计学等诸多学科,算法管理模块除了提供常见的数据挖掘算法节点,比如决策树、神经网络、生存分析等,还提供算法接口,用户可以自己实现算法,然后添加到平台中,就可象使用平台提供的常见算法一样和其它节点(数据挖掘算法节点、数据预处理节点)一起配合使用,组成一个可视化的、完整的数据挖掘流程。

模型管理模块对数据挖掘算法生成的模型进行管理,主要提供两方面的功能:

第一、将每个模型都包装成节点,并摆放在指定的位置,用户可以像使用其他节点一样,用可视化的方式使用这些模型对数据执行预测、分析等操作,并产生输出结果。

第二、实现对生成的模型进行增、删、改、查等编辑维护工作。

### 3.4 商业逻辑管理

平台通过将数据挖掘、商业逻辑、行业应用相结合,对各行业常见的商业逻辑进行抽象,抽取其共性的内容,定义成通用的商业逻辑节点供用户使用。此外,用户也能够在可视化的界面下以接近自然语言的方式定义自己需要查找的内容,并将定义结果以模型形式存储,在平台运行时表现为一个模型节点,和其它节点一起供用户使用,用户可对定义内容进行维护,从而降低用户对专业知识的要求,为推进数据挖掘在各行业中的广泛应用提供了可能性。

### 3.5 模型评价管理

用户能够根据分析的目的灵活选择多种数据挖掘模型,并产生对应的分析结果。模型应用子系统自动评价这些分析结果,对用户的问题,提出建设性的意见,帮助用户确定最切合实际的分析模型。这个子系统虽然功能比较少,但是实现的难度是几个子系统中最大的。

## 4 平台实现及技术特点

平台容易使用和维护往往意味着实现上的复杂,平台采用了多种先进的技术,多角度、多层面地保证平台的质量特性得以实现。具体如下:

在技术选型的时候,我们重点考虑所用技术的先进性、成熟性、使用广泛行,有些技术可能更先进,但是,考虑到目前在数据挖掘工具开发中使用的广泛程度和技术成熟度,我们最终还是没有选择。

在系统设计的时候,突出考虑平台易使用性、可扩展性、易维护性、可靠性等质量特性。

在平台开发工具选择上,我们决定采用 DELPHI 7.0 + ODBC + Microsoft SQL Server 2000 来开发平台,因为 DELPHI 7.0 面向对象和所见即所得特性、ODBC 开放性以及 Microsoft SQL Server 2000 强大的数据管理能力对于提升平台的易使用性、可扩展性、易维护性有重要作用。

在平台架构设计上，混合采用三层 C/S（客户机/服务器）和 B/S（浏览器/服务器模式），通过并行和负载平衡来提高系统的执行效率，数据预处理、数据挖掘等操作采用就近原则，将相关模块靠近数据库部署，这部分的功能模块采用三层 C/S 方式实现，而挖掘结果、数据信息、内容查询部分的功能模块则采用 B/S 的方式实现，把开发、维护等几乎所有工作都集中在服务器端，当需要对网络应用进行升级时，只需更新服务器端的软件就可以，无须在客户端进行平台的升级和维护，这样有利于提高系统可扩展性和易维护性，有利于降低系统维护与升级的成本。

在后台数据库设计上，将不同类型的数据分类集中在不同类型的数据库中，并对不同类型数据库根据不同用途采用不同的指导原则。因为元数据库的增、删、改比较频繁，关系比较复杂，故采用范式理论来指导数据库的设计，要求所有表要达到 3NF 或 BCNF 的标准。而对数据挖掘库中的数据，由于主要是进行各种复杂的查询，查询涉及数据量大、牵涉面广，故采用反范式作为设计指导，按照数据仓库的要求组织数据，通过适当增加冗余，减少表之间的连接，提高查询效率。

## 5　结束语

本文介绍了数据挖掘商业应用平台各个模块的功能，探讨了平台分析设计过程中要考虑的内容，给出了平台分析与设计方案，方案综合采用多种先进理论、技术和工具，并最大限度发挥各种技术的优势，以确保能够高质量地实现平台，并且实现的平台能满足各种用户的需要，能降低对用户的专业知识要求。该应用平台已设计完毕，并经用户使用，表明该应用平台是有效的，可以大幅度提高数据挖掘效率，且对非专业用户很容易使用，这对于在企业进一步推广数据挖掘应用是有相当的现实意义及实用价值。

## 参考文献

1　刘绍清、黄章树，数据挖掘商业应用平台的需求分析，待发表.

2　刘绍清、黄章树、黄剑辉，数据挖掘商业应用平台的数据预处理管理[J]，重庆工商大学学报（自然科学版），2006 年第 5 期，453－456.

3　（美）W·H·Inmon. 王志海，林友芳译. Building the Data Warehouse [M]，北京：机械工业出版社，2003，3－150.

4　李珊珊、陈维斌，数据仓库中元数据标准的对比研究[J]，福建电脑，2005 年 10 期，52－53.

5　叶琪，决策支持系统中的数据预处理[J]，微型电脑应用，2003 年第 19 卷第 11 期，46－47.

6　朱明，数据挖掘[M]，合肥：中国科技大学出版社，2002 年，37－50.