数据挖掘商业应用平台的需求分析

Requirements Analysis of Data Mining Platform for Commercial Use

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

摘要：本文论述我们为广大企业用户开发的数据挖掘商业应用平台的需求分析，对数据挖掘的业务流程和数据流程进行了详尽的分析，并给出了数据流程图。重点探讨了平台应该具备的功能需求和性能需求，并给出了详细的需求描述。

关键词：数据挖掘 商业应用平台 需求分析 业务流程 数据流程

本文介绍一个通用的数据挖掘辅助工具：数据挖掘商业应用平台，本文重点从需求分析的层面探讨为了让使用者不需要面对复杂的算法，不需要掌握众多非常专业的数据挖掘工具及高超的技巧，同时又能够保证数据挖掘质量，该平台应该具备的功能需求和性能需求，只有这样，才能有效指导系统的设计和实现，保证系统的设计目标得到实现。

1 数据挖掘业务流程分析

数据挖掘是一个多阶段的、复杂的、高难度的系统工程，在实施数据挖掘项目之前，制定一个详细计划能保证数据挖掘有条不紊的实施并取得成功。CRISP－DM（跨行业数据挖掘标准流程）是最主要数据挖掘过程模型之一，它从方法学的角度强调实施数据挖掘项目的方法和步骤，它把数据挖掘业务流程分为以下六个阶段[2]：

1.1 业务理解（Business Understanding）

这个阶段是整个数据挖掘过程中最重要的阶段，它重点从商业角度去理解项目目标和客户真正的需求，进而把这些理解转化为一个数据挖掘的定义，并进而设计出一个达到目标的初步方案，能否正确理解业务，定义合理的、可行的商业问题决定着数据挖掘项目的成败。

1.2 数据理解（Data Understanding）

数据理解阶段开始于数据的收集工作，然后是熟悉数据的工作，具体如：检测数据的质量，对数据有初步的理解，探测数据中比较有趣的数据子集，进而形成对潜在信息的假设等。

1.3 数据准备（Data Preparation）

数据准备阶段涵盖了从原始数据中构建挖掘用数据集的全部工作。作为数据挖掘对象来源的原始数据，常常包含着噪声、不完整、甚至是不一致的数据。为了得到高质量的数据挖掘效果，在进行数据挖掘之前，必须对原始数据做一定的预处理，它是整个数据挖掘过程中一个重要步骤[1]。它包括数据清洗、集成、转换和消减[3]四个方面的内容。

1.4 建立模型（Modeling）

这是整个 CRISP－DM 的关键阶段。选择合适的数据挖掘技术，并将其参数不断调整，直到最优值。由于建模需要，必要的时候可以退回到数据准备阶段。

1.5 模型评估（Evaluation）

模型构建后，通过对模型评价指标值的判断、对验证数据集预测效果的比较和企业真实数据预测效果的评价，最终得出模型的评估是非常重要和必要的。

1.6 结果发布（Deployment）

模型的创建并不是项目的最终目的，建模的目的是为了增加更多有关于数据的信息，但这些信息仍然需要以一种客户能够使用的方式被组织和呈现。

2 数据流程分析和数据流程图

在详细介绍 CRISP－DM 定义的数据挖掘业务流程的基础上，绘制出数据挖掘的数据流程图，其顶层和

第 0 层的数据流程图如图 1 和图 2 所示。

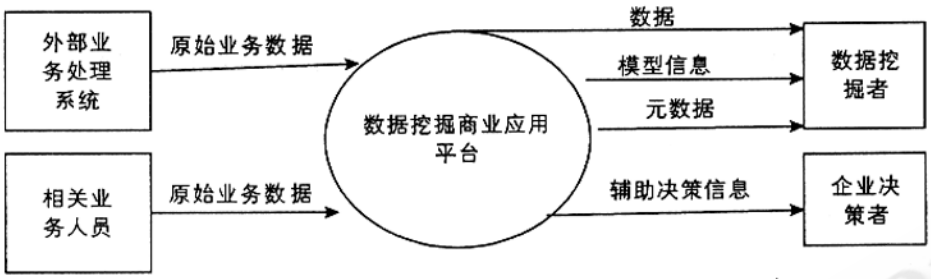

图 1　顶层数据流程图

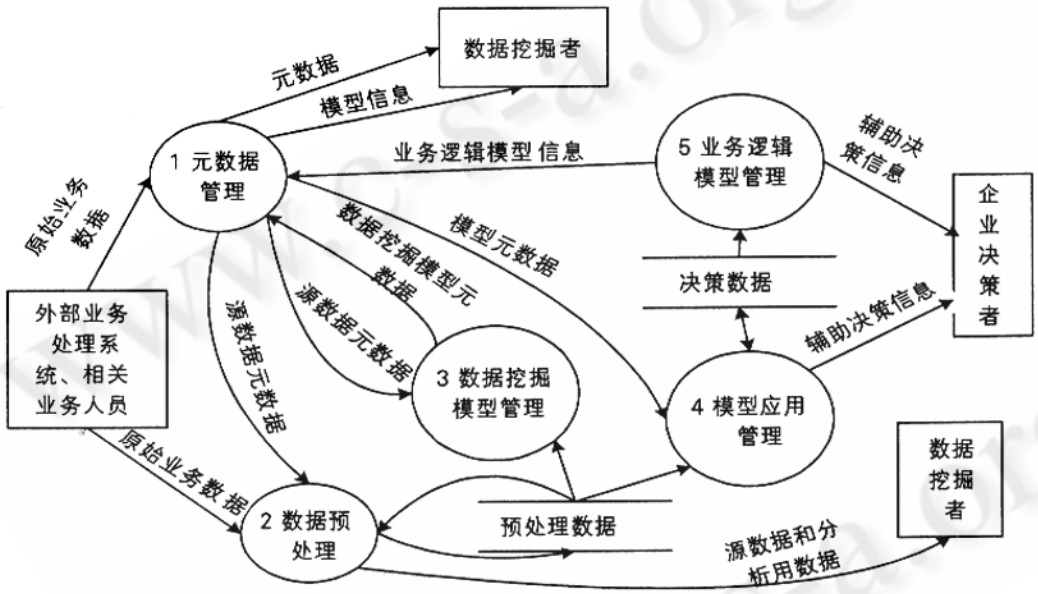

图 2　第 0 层数据流程图

3　功能需求分析

通过前面论述的平台开发目标、数据挖掘业务流程和数据流程特点和内容要求的论述，可以得出，平台应该具备以下功能：

3.1　全面的元数据管理功能

因为数据挖掘系统所面对的数据来源是多样化的、分散的、内容经常是不一致的，所以，平台必须提供元数据管理的功能，对元数据从生成到应用到退出的整个生命周期进行管理，保证企业数据定义的一致性、数据间关系清晰明确，从而方便用户访问、使用和沟通，使业务用户、业务分析人员、数据分析人员和开发人员清楚、全面、一致地了解企业数据资产的信息。具体包括以下功能：

(1) 元数据的生成。元数据的来源主要两种，一种是描述数据源的信息；另外一种是描述挖掘处理过程的信息，比如数据预处理流程信息、数据挖掘节点信息、数据转换规则信息等，它在数据挖掘过程中逐步产生。

(2) 为元数据添加注释说明。为给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，应能允许用户对所收集的元数据添加注释，并显示相关信息。

(3) 元数据引退。为了缩减元数据规模，降低元数据管理的难度，提高元数据的管理和处理效率，对于一些已经过时或暂时不用的元数据信息应该予以引退，引退包括两种：删除和导出。删除操作直接将元数据删除而不可恢复，导出操作则将元数据备份到指定设备，是可恢复的。

3.2　轻松地理解数据

拥有丰富数据是数据挖掘项目得以进行的基础，而数据的准确性则是数据挖掘项目成功的保证，但是二者往往是互相矛盾的，丰富的数据往往会加大正确理解数据的难度，平台应该允许使用者以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；帮助使用者从全局和局部的视角全面观察数据，了解原始数据如何经过一步一步的处理达到当前的状态。这是平台要解决的一个重点。

3.3　数据取样

数据挖掘往往是针对海量数据进行的，对数据的任何一个操作都可能需要耗费很长的时间。而数据挖掘的特点却又决定了很多工作需要多次反复的探索，比如，数据理解、数据抽取、转换、载入、建模等，这很可能就意味着数据挖掘过程将在不断的等待中度过，其效率之低是可以想象的。

为了提高数据挖掘效率，特别是在初始阶段对数据结构和内容的探索与理解过程中减少不必要的等待时间，数据挖掘商业应用平台应该提供数据取样的功能，从数据源抽取小部分的数据样本，对样本的大部分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。

3.4 数据抽取、转换和载入(ETL)

平台应该允许用户根据数据挖掘的目的，数据本身的特点和不同数据挖掘算法对数据的需求，有机地组合各种预处理方法，用可视化地方式进行清洗和整理，从而大大地简化数据挖掘的数据准备过程。具体功能如下：

(1) 数据属性统一规范命名。数据挖掘经常涉及多个数据库多张表，属性命名往往缺乏统一标准，随意性比较大，突出的表现有：同义异名、同名异义、不规范命名、名义不符等[2]。ETL 应该能够对不同表的属性根据其含义重新定义其在数据挖掘库中的名字，并以转换规则的形式存放在元数据库中，在数据集成的时候，平台自动根据这些转换规则，将源数据中的字段名转换成新定义的字段名，从而实现数据挖掘库中的数据属性规范命名，名义相符，并且同名同义。

(2) 实现数据缩减，大幅度缩小数据量。ETL 应该允许用户采用数据聚合、数据压缩、数据块消减、维度缩减、主成分分析、属性类型转换、内容转换等方法，在不影响(或基本不影响)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。

(3) 实现数据清洗和集成。平台应该提供清洗功能，消除储存数据中存在的错误、异常(偏离期望值)、冗余、冲突以及数据内涵不一致(如：部门编码在不同表中出现不同值)等现象，最终按照统一的格式结合在一起形成比较完整的数据集合，为数据挖掘提供数据基础。

3.5 挖掘结果的展示功能

如何对数据挖掘结果进行评估，较好地解决模式的易懂性问题是平台要解决的一个重要的问题[4]。数据经模型处理后的结果必须以多种展示方式(包括语言、图形化界面等)提供给用户，还应提供灵活的展示操作，例如针对立方体的切片，旋转等。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，也可以是 Excel 文件形式以方便进一步的加工处理。

3.6 商业逻辑解释功能

由于数据挖掘商业应用平台的最终用户非专业用户，所以，平台应该能够提供可视化界面，让用户用他们熟悉的商业逻辑来表达挖掘需求，平台自动将这些挖掘任务转换成具体的商业模型，从而将数据挖掘、商业逻辑、行业应用相结合，降低平台对用户使用的要求。

3.7 提供多种数据挖掘模型生成能力

平台除了提供目前在数据挖掘实践中常见的数据挖掘算法，如：统计方法、关联分析、聚类分析、神经网络、决策树等之外，还应该能够提供用户自定义数据挖掘算法的功能，让这些算法和平台提供的算法一起无缝工作。

3.8 提供单一流程实现数据挖掘的整个过程的能力

平台应该能够允许用户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，这是本平台的一个重要特点。

4 性能需求分析

4.1 响应时间需求分析

该平台需要面对两种不同类型的用户：数据挖掘者和挖掘结果使用者。二者关心的内容不同，对数据操作要求有很大的不同，直接导致其对平台性能要求也不同，平台在设计的时候应该充分考虑这点。

数据挖掘者一般都是对海量的数据进行复杂的操作，不同的操作内容、操作顺序以及数据集合大小对响应时间的影响很大，无法定量规定，所以在平台按照指令对数据操作的过程中，平台应该能够明确指示处理

的进度,预估剩余处理时间。对样本数据进行操作以验证指令正确性和了解数据结构的时候,单步操作的响应时间和运行时间总共应该在 2 秒以内,整个流程的响应时间应该在 20 秒以内。

挖掘结果使用者很可能不是计算机专家,且在地域上比较分散,他们主要关心的是数据挖掘的结果,而不是数据挖掘处理过程。此时,对平台的响应时间要求就比较高,对平台的易使用性要求也比较高。

4.2 易使用性需求分析

平台的一个主要特点就是能降低对使用者的要求,非专业用户能轻松使用,因此,对平台的易使用性提出了很高的要求。平台应该提供可视化的界面,将具体算法计算过程、数据操作过程的详细内容包装起来,使用者用可以用类似自然语言或他们熟悉的语言定义操作指令,而不需要了解具体操作的详细实现内容,不需要接触底层的算法调用。

4.3 可扩展性需求分析

平台应该保持一定可扩展性,主要体现在以下两方面:

第一、能够让用户自主地添加新的数据挖掘算法,而不需要重新编译整个平台。

第二、能够让用户定义一些基本或常用的数据挖掘步骤,完成数据挖掘项目的时候,通过组合这些步骤来快速建立一个数据挖掘过程。

5 系统设计

在平台模块功能设计上,结合前面需求分析、数据挖掘的特点和我们的目的要求,整个平台从功能上可以分成以下 5 个子系统:元数据管理子系统分为数据源连接、取回数据样本、抽取元数据、元数据编辑四个模块;数据预处理子系统实现数据属性统一规范命名、数据缩减、数据清洗和集成;数据挖掘管理子系统允许用户添加自己定义的数据挖掘算法;商业逻辑管理子系统使用户能够以接近自然语言的方式定义自己需要查找的内容,从而降低用户对专业知识的要求;模型应用子系统可对模型进行分析和评价,确定最切合实际的分析模型。

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),数据预处理、数据挖掘等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、内容的查询部分的功能模块则采用 B/S 的方式实现,以充分发挥二者的优点。

在后台数据库设计上,将后台数据库分为元数据库和数据挖掘库,采用数据范式理论,分别进行规范化设计和反范式设计。对增、删、改比较频繁,关系比较复杂的元数据库,采用规范化设计,而对主要进行各种复杂的查询,查询涉及数据量大的而数据挖掘库,采用反范式设计,通过适当增加冗余,减少表之间的连接,提高查询效率。

6 结束语

认真做好平台的需求分析对平台的开发有非常重大的意义的,本文通过对数据挖掘业务流程、数据流程以及相关背景的分析和理解,可知道平台需要“做什么”,为后续的设计与实现提供指导依据,保证系统的开发质量和效率,确保平台达到适合非专业用户,满足不同类型用户的需求,提高管理水平的设计目标,关于本应用平台的设计与实现将另文介绍。

参考资料

1 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.

2 潘无名、潘云鹤, 数据挖掘过程的多维视图[J], 计算机应用研究, 2004 年 08 期,211 –216.

3 朱明, 数据挖掘[M], 合肥:中国科技大学出版社, 2002 年,1 –50.

4 徐燕、柳长安、祖向荣, 基于虚拟现实技术的数据挖掘结果可视化[J], 计算机应用研究,2004 年第 21 卷第 12 期,190 –192.