

• 软件与算法 •

基于 MVC 架构的数据挖掘平台的设计与实现

叶苏南, 彭 宏, 覃姜维

(华南理工大学 计算机科学与工程学院, 广东 广州 510006)

摘 要: 为了增强数据挖掘软件各功能模块的可扩展性和复用性, 分析了现有数据挖掘工具的优缺点, 并综合考虑数据挖掘过程的实际特点, 提出了一个基于 MVC 架构的数据挖掘平台设计方案。在此基础上, 利用 Eclipse plug-in, RCP, GEF 等技术, 实现了一个数据挖掘平台原型系统。该平台遵循 CRISP-DM 过程标准, 在软件架构上实现了低耦合、高复用, 为用户提供了一个友好、灵活、易重用、可扩展的数据挖掘应用环境。

关键词: 数据挖掘; 软件复用; 软件构件; Eclipse 插件; 富客户端平台; 图形编辑框架; 模型-视图-控制器

中图分类号: TP311; TP391 **文献标识码:** A **文章编号:** 1000-7024 (2010) 05-1013-04

Design and implementation of data mining platform based on MVC structure

YE Su-nan, PENG Hong, QIN Jiang-wei

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: To increase expansibility and reuse of functional modules of data mining software, the advantages and disadvantages of existing data mining tools are analyzed, and characters of data mining process are also taken into consideration, a design scheme of data mining platform based on MVC architecture is presented. Based on these, a data mining platform of prototype system is implemented by Eclipse plug-in, RCP and GEF technologies. The platform according to CRISP-DM implements loosely-coupled and reusable software architecture and provides a friendly, flexible, easy reuse and extensible environment of data mining application.

Key words: data mining; software reuse; software component; Eclipse plug-in; RCP; GEF; MVC

0 引言

由于各行各业均积累了海量的数据, 这些数据中通常都蕴涵着丰富的有价值的知识, 借助数据挖掘技术可以充分发掘出这些知识, 为企业的分析决策者起到很好的辅助支持作用。因此, 数据挖掘一直是众多学者的研究热点。在国外, 数据挖掘在金融业、保险业、零售业、生物医学等领域已经有了广泛的应用。在国内, 在数据挖掘技术的理论研究上已经取得了许多成果, 但在数据挖掘软件方面的研究则刚刚起步。

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中, 提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程^[1], 数据挖掘是知识发现过程的一个关键步骤。知识发现过程是一个高级的、复杂的处理过程, 通常需要借助数据挖掘软件去完成各种任务。目前已经有不少的数据挖掘软件被开发出来, 例如 IBM Intelligent Miner、SAS Enterprise Miner、SPSS Clementine、Weka 等。这些软件基本都实现了多种经典的数据挖掘算法, 但在扩展性和易用性方面都存在着一些问题, 例如商业数据挖掘软件, 出于竞争等各方面因素, 通常难以对其进行功能扩展和模块的重用,

用户无法根据自己的需要添加新的算法, 只能使用软件提供的算法; 而一些开源数据挖掘软件, 例如 Weka 虽然提供算法接口, 支持用户自己添加新算法, 但却不能够对结果可视化进行扩展, 并且在界面友好程度, 易用性方面都有所欠缺。鉴于这些问题, 如何设计一个低耦合、高复用、扩展性强且方便使用的数据挖掘平台是非常具有现实意义的, 本文介绍了我们设计并实现的一个数据挖掘平台, 该平台具有友好的用户界面, 其各功能模块之间实现松散耦合, 便于用户进行功能的扩展和模块的重用, 方便二次开发。

1 基本概念

1.1 数据挖掘系统的结构

数据挖掘系统按照其应用的不同可分为通用型数据挖掘系统和面向特定领域数据挖掘系统两类^[2]。数据挖掘是从存放在数据库、数据仓库或其它信息库中的大量数据中发现有趣知识的过程, 基于这种观点, 一个典型的数据挖掘系统如图 1 所示^[3]。

1.2 数据挖掘过程模型

不同的研究机构和组织, 对数据挖掘过程的划分是略有

收稿日期: 2009-05-15; 修订日期: 2009-07-23。

基金项目: 广东省自然科学基金项目 (07006474); 广州市科技攻关基金项目 (2007B010200044)。

作者简介: 叶苏南 (1984—), 男, 福建周宁人, 硕士研究生, 研究方向为数据挖掘、机器学习; 彭宏 (1956—), 男, 重庆人, 博士后, 教授, 研究方向为人工智能应用技术、智能商务与数据挖掘、智能网络技术等; 覃姜维 (1984—), 广西河池人, 博士研究生, 研究方向为可视化数据挖掘、机器学习。E-mail: zhazha1984@gmail.com

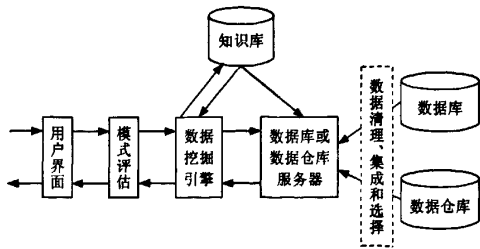


图1 典型的数据挖掘系统结构

不同的。目前公认的,较有影响力的数据挖掘过程模型是CRISP-DM(cross-industry standard process for data mining),即为“跨行业数据挖掘过程标准”^[4]。此KDD(knowledge discovery in data-bases)过程模型于1999年欧盟机构联合起草,通过近几年的发展,在各种KDD过程模型中占据领先地位,采用量达到近60%^[5]。CRISP-DM将数据挖掘过程划分为6个步骤,即商业理解、数据理解、数据准备、建模、评估、部署。CRISP-DM模型如图2所示。

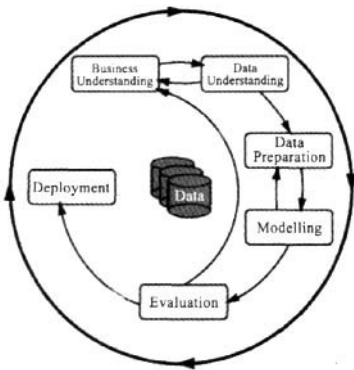


图2 CRISP-DM过程模型

1.3 MVC架构

MVC(model-view-controller,模型—视图—控制器)用于表示一种软件架构的设计模式。MVC模式起源于Smalltalk,它把软件系统分为3个基本部分:模型,视图和控制器^[6]。MVC作为一种模块化设计思想,可以降低软件系统各个模块间的耦合性,提高程序代码的可重用性,并且便于软件系统的后期维护以及功能扩展。MVC的架构如图3所示。

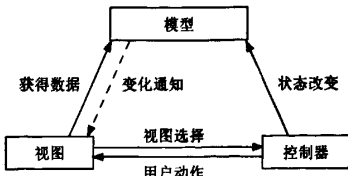


图3 MVC架构

2 数据挖掘平台的设计与实现

本文设计的数据挖掘平台主要由MVC模块、图形用户界

面、数据挖掘构件库、构件管理模块4个功能模块组成。下面分别对系统的各个功能模块的设计及实现展开说明,整个系统结构如图4所示。

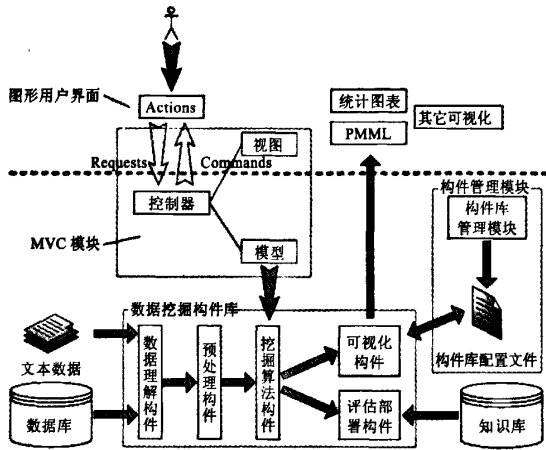


图4 数据挖掘平台系统结构

2.1 MVC模块

MVC模块是数据挖掘平台的核心,主要利用GEF(graphical editor framework)图形编辑框架来实现,GEF提供了标准的MVC结构,它允许开发人员以图形化的方式展示和编辑模型,从而提升用户体验。下面就MVC结构在数据挖掘平台中的应用展开说明。

(1)模型:模型用于封装与数据挖掘业务相关的数据及其操作,系统以构件的形式来管理这些各不相同的业务数据及其操作方法。模型不依赖视图和控制器,它业务处理流程相对于其它层来说是黑箱操作,用户在视图上的各种操作请求会被发送给控制器,由控制器根据请求类型通知模型调用相应的操作方法,并根据最终的处理结果刷新视图。

(2)视图:视图是以可视化的形式将模型的信息展现出来。数据挖掘平台包含了挖掘流程设计器、大纲视图和属性视图。挖掘流程设计器提供了一个编辑区域,在该区域内,用户可以根据业务需要对构件拖拽连接快速地建立起一个数据挖掘流程,并且可以对每个构件进行管理,例如参数设置、运行构件、结果查看等操作。大纲视图和属性视图则是对挖掘流程设计视图中的内容进行辅助显示。大纲视图以列表和缩略图的形式对当前流程设计视图中的所有构件进行快速预览和定位,而属性视图负责显示各个构件的一些信息。

(3)控制器:在MVC模块中,控制器是模型和视图之间的唯一桥梁。控制器相当于一个分发器,它从视图处接收用户的操作请求,然后将这些操作请求发送到对应的模型,执行相应的操作,最后将结果反馈到视图上。控制器并不做任何的数据处理,它只是通知模型,真正的数据处理则是在模型中完成的。模型、视图与控制器的分离,使得一个模型可以具有多个显示视图。如果用户通过某个视图的控制器改变了模型的数据,所有其它依赖于这些数据的视图都应反映到这些变化。因此,无论何时发生了何种数据的变化,控制器都会将变化通知到所有相关的视图,导致其刷新。

2.2 图形用户界面

图形用户界面采用 Eclipse RCP(eclipse rich client platform)富客户端平台开发,因此可以直接继承Eclipse的风格与功能,极大地加快了开发速度和避免重复性工作。图形用户界面主要是提供数据挖掘流程的设计及其相关的管理操作功能,如图5所示。

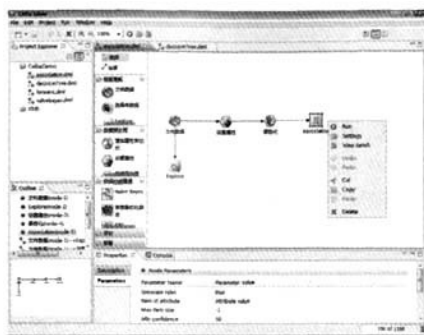


图5 数据挖掘平台原型系统

扩展点(extension point)是 Eclipse plug-in 和 RCP 开发中的一个重要机制,通过扩展点可以向数据挖掘平台添加新的功能,例如在MVC模块中介绍的挖掘流程设计器、大纲视图、属性视图都是利用扩展机制添加到数据挖掘平台中的。以工程管理功能为例,扩展点的配置信息如下:

```
<extension point="org.eclipse.ui.views">
    <view
        id="cn.edu.scut.mfdm.ui.CaseNavigator"
        name="Project Navigator"
        category="cn.edu.scut.mfdm.navigatorcategory"
        class="org.eclipse.ui.navigator.CommonNavigator"
        allowMultiple="false">
    </view>
    ...
</extension>
```

例子中的扩展点是直接使用 Eclipse 的工程管理能力,扩展点也可以对 Eclipse 功能进行扩充,例如数据挖掘平台中的大纲视图和属性视图,甚至可以根据需要定义新的功能。扩展点是相对独立的,添加或删除某个扩展点并不会影响到其它的扩展点,因此,扩展点的工作机制非常方便用户在图形界面上进行二次开发。

2.3 数据挖掘构件库

软件复用是在软件开发中避免重复劳动的解决方案。通过软件复用,可以提高软件开发的效率和质量。软件复用被视为解决软件危机、提高软件生产效率和质量的现实可行的途径^[7]。软件复用从最初的子系统调用开始,经历了结构化方法和面向对象方法等几次重大的发展,逐步形成了通过建立类库、构件库、框架库和模式库等多种方法^[8]。其中,构件库是实现软件复用的重要依托,软件复用的成功与否在很大程度上取决于构件库的结构、成分、管理方式等。结合 CRISP-DM 过程模型,通过对数据挖掘流程的各阶段操作进行构件化,避免

开发数据挖掘软件过程中的重复工作,提高数据挖掘软件开发的效率和质量,还可以在一定程度上提高数据挖掘的质量。数据挖掘构件库的层次结构如图6所示。

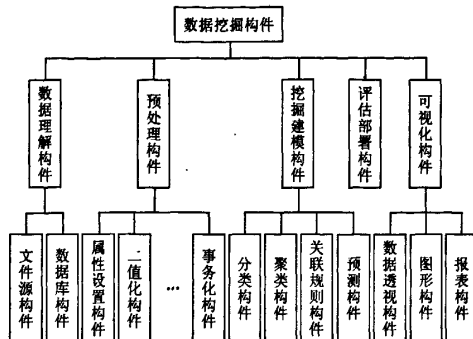


图6 数据挖掘构件库层次结构

数据挖掘平台将整个业务划分为数据理解、预处理、挖掘建模、评估部署、可视化5个大构件,这些大构件又进一步细分为若干个小的业务构件。各业务构件之间是相互独立的,是在数据、操作上相对封闭的一个完整子集。各业务构件对外提供统一的访问接口,在设计上做到相互之间的无关性,在添加更改业务构件时,原有业务构件系统不会受到任何影响。通过这种分层式的划分,使整个构件库具有良好的灵活性、扩展性、可维护性和开放性。

2.4 构件管理模块

构件库管理模块对构件库进行统一的管理和控制。利用 Eclipse plug-in 的插件更新机制,构件管理模块还支持在线更新现有构件或者添加新构件。此外,构件库管理系统还承担着构件库的维护工作,保证构件库的安全性和完整性。构件管理模块通过构件库配置文件实现对数据挖掘构件库的管理,数据挖掘平台在启动时根据配置文件的内容将构件加载到挖掘流程设计器左侧的构件工具箱中。构件库配置文件的内容如下:

```
<Component>
    <Name value="文件数据"/>
    <Description value="从文件中输入数据"/>
    <Icon value="icons/nodes/InputDataFile.png"/>
    <OperatorClass value="scut.mfdm.node.operator.InputDataFileOperator"/>
    <SettingsClass value="scut.mfdm.node.settings.InputFileDataSettings"/>
    <ResultClass value="scut.mfdm.node.result.DataOperatorResult"/>
</Component>
```

构件库配置文件中记录了构件的名称,描述和显示信息,以及构件的操作、参数设置和结果展示3个Java类,每个构件都是通过这3个类描述自己的业务数据和操作方法。由于构件库管理模块是利用Java的类反射机制来动态创建3个Java类的具体对象,因此在二次开发添加新构件时,只需根据业务需要实现这3个类或利用挖掘平台中现有的类,然后修改构

件库配置文件,至于数据挖掘平台是如何调用这些对象则无需考虑,因此,极大地提高了软件的可扩展性和灵活性。

3 数据挖掘平台的应用

本数据挖掘平台采用开发式架构,为算法应用提供了一个良好的环境。该平台为数据源,算法,可视化等各类构件的编写订立了良好的规范,只要符合此规范,即可以加入到平台中来,与其它组件相互作用。并且,还着重考虑如何让用户使用更方便更容易,利用了 RCP 和 GEF 技术,极大地提高了用户的体验。目前,该平台已经支持多种数据源,实现了多种数据预处理算法和经典的数据挖掘算法,并且在数据挖掘可视化方面也做了一定的工作,可以用于金融业、保险业、零售业、科学研究、工业部门、司法部门、生物医学、网上电子商务的商品信息等领域的搜索、分析和决策,为各级经营决策者提供有效的决策支持和信息服务,有着广阔的应用前景。

4 结束语

论文针对当前数据挖掘软件结构上存在的难以重用,扩展的问题,提出了一种基于 MVC 架构的开放式数据挖掘平台,详细介绍了其系统架构的设计以及各部分的功能。该平台的各功能模块在结构上通过松散耦合相互关联,而图形用户界面和数据挖掘构件库也都满足了软件开发的“开闭原

则”。因此,无论是算法研究和开发人员,或是数据决策分析人员,又或是二次开发应用人员都可以通过该数据挖掘平台的得到一个灵活、一体化的解决方案。

参考文献:

- [1] 邵峰晶,于忠清.数据挖掘原理与算法[M].北京:中国水利水电出版社,2003.
- [2] 陆晶,赛英.基于 C/S 体系结构的数据挖掘平台的设计[J].计算机工程与设计,2005,26(3):598-600.
- [3] Jiawei Han,Micheline Kamber.数据挖掘概念与技术[M].北京:机械工业出版社,2006.
- [4] Lukasz A Kurgan, Petr Musilek. A survey of knowledge discovery and data mining process models[J].Knowledge Engineering Review,2006,21(1):1-24.
- [5] Cios K, Kurgan L. Trends in data mining and knowledge discovery[M].London:Springer,2005.
- [6] Brown D,Davis CM,Stanlick S.Struts2 in action[M].Greenwich, CT:Manning Publications Co,2008.
- [7] 张翔,周明全,耿国华.软件复用与基于 Java 的 COM 组件实现[J].计算机应用与软件,2003,20(7):80-82.
- [8] 唐勇敏.以构件为核心的软件工业化的生产方式[J].计算机应用,2006,26(12):225-227.

(上接第 978 页)

```
exec sp_executesql
```

```
N'select EmployeeID from Employees where FirstName = @
FirstName and LastName = @LastName',N'@FirstName nvarchar
(4000),@LastName nvarchar(4000)',@FirstName=N''' or 1=1--',@
LastName=N''
```

从以上代码可以看出,参数化查询与调用一个存储过程是非常相似的,因为它处理的输入是作为参数在使用,而不是字符串。

通过优化与整合代码,建立一个具有很多 where 条件的存储过程,可以有效地防范在一个参数化查询时的 SQL 注入。因为参数化查询要求用户带权限才可以访问特定的系统表和视图,这种方法在安全数据查询时并不被推荐,而存储过程在被执行时,因为其不需要具有访问任何表或者视图的权限,所以存储过程在防范一个参数化查询的 SQL 注入是非常实用的。

4 结束语


本文对 SQL 注入攻击的方法、原理以及常用注入途径进行了描述和总结,并着重阐述了使用输入验证、SQL Server 防御和使用存储过程替代参数化查询相结合的方法来构建防范 SQL 注入执行模块的思路和方法。对于普通用户误操作和低等级恶意攻击,客户端检测能够自动将其屏蔽;考虑到客户端检测有可能被有经验的攻击者绕开,特在服务器端设定二级检测。在文中还提出了对高等级恶意攻击的自动备案技术,

并给出了相应代码。本文对于越来越多的 Web 动态应用程序提供了较好的防范 SQL 注入式攻击的解决方案,具有一定的现实意义。

参考文献:

- [1] 赛门铁克第 2 期.Symantec 资讯[EB/OL].http://www.cww.net.cn/zhuanti/html/2009/2/10/200921095084992.htm.
- [2] Anley C.Advanced SQL injection in SQL server Applications [EB/OL].http://www.creangel.com/papers/advanced_sql_injection.pdf, An NGSSoftware Insight Security Research (NISR) Publication,2002.
- [3] 陈小兵,张汉煜,骆力明,等. SQL 注入攻击及其防范检测技术研究[J].计算机工程与应用,2007,43(11):150-152.
- [4] Cerrudo C.Manipulating Microsoft SQL server using SQL injection [EB/OL].http://injection.rulezz.ru/Manipulating_SQL_Server_Using_SQL_Injection.pdf.
- [5] 赖滇,黄宇.一种缓冲区溢出漏洞分析与探测算法 BOVADA[J].计算机工程,2006,27(18):52-53.
- [6] 谷震离,杜根远. SQLServer 数据库应用程序中数据库安全性研究[J].计算机工程与设计,2007,28(15):3717-3719.
- [7] 池瑞楠.Windows 缓冲区溢出攻击的实例研究[J].微计算机信息,2007(3):39-40.
- [8] 黄景文.SQL 注入攻击的一个新的防范策略[J].微计算机信息,2008(6):31-32.

基于MVC架构的数据挖掘平台的设计与实现

作者: [叶苏南](#), [彭宏](#), [覃姜维](#), [YE Su-nan](#), [PENG Hong](#), [QIN Jiang-wei](#)
作者单位: [华南理工大学计算机科学与工程学院, 广东, 广州, 510006](#)
刊名: [计算机工程与设计](#) 
英文刊名: [COMPUTER ENGINEERING AND DESIGN](#)
年, 卷(期): 2010, 31(5)
被引用次数: 7次

参考文献(8条)

1. 邵峰晶;于忠清 [数据挖掘原理与算法](#) 2003
2. 陆晶, 赛英 [基于C/S体系结构的数据挖掘平台的设计](#) [期刊论文]-[计算机工程与设计](#) 2005(3)
3. Jiawei Han;Micheline Kamber [数据挖掘概念与技术](#) 2006
4. LUKASZ A. KURGAN;PETR MUSILEK [A survey of Knowledge Discovery and Data Mining process models](#) [外文期刊] 2006(1)
5. Cios K;Kurgan L [Trends in data mining and knowledge discovery](#) 2005
6. Brown D;Davis CM;Stanlick S [Struts2 in action](#) 2008
7. 张翔, 周明全, 耿国华 [软件复用与基于Java的COM组件实现](#) [期刊论文]-[计算机应用与软件](#) 2003(7)
8. 唐勇敏 [以构件为核心的软件工业化的生产方式](#) [期刊论文]-[计算机应用](#) 2006(z2)

本文读者也读过(2条)

1. 刘亮, 霍剑青, 郭玉刚, 袁泉, 王晓蒲, LIU Liang, HUO Jianqing, GUO Yugang, YUAN Quan, WANG Xiaopu [基于MVC的通用型模式的设计与实现](#) [期刊论文]-[中国科学技术大学学报](#) 2010, 40(6)
2. 郭俊荣, 李洁, GUO Jun-rong, LI-jie [MVC在Eclipse RCP开发中的应用](#) [期刊论文]-[煤炭技术](#) 2010, 29(8)

引证文献(7条)

1. 李志奎, 丁立群, 关英宇 [一体化缴费接入管理平台的数据架构设计与优化](#) [期刊论文]-[网络安全技术与应用](#) 2012(04)
2. 王青峰, 翟永刚, 林楠 [通用数据挖掘平台设计与实现](#) [期刊论文]-[信息通信](#) 2012(02)
3. 侯瑞春, 胡青霞, 丁香乾, 周连荣 [基于GEF的业务模式图形编辑器的设计与实现](#) [期刊论文]-[现代电子技术](#) 2012(20)
4. 顾宝潮 [基于.Net平台的病理信息系统的设计与实现](#) [学位论文] 硕士 2011
5. 朱诗生, 王正超, 周明明, 陈晓强 [基于插件技术的VWESA平台的研究与设计](#) [期刊论文]-[计算机应用](#) 2012(z1)
6. 吴志伟 [光缆网络运营分析系统](#) [学位论文] 硕士 2010
7. 王正超 [基于插件技术的ERP系统开发平台的研究与设计](#) [学位论文] 硕士 2012

引用本文格式: [叶苏南](#), [彭宏](#), [覃姜维](#), [YE Su-nan](#), [PENG Hong](#), [QIN Jiang-wei](#) [基于MVC架构的数据挖掘平台的设计与实现](#) [期刊论文]-[计算机工程与设计](#) 2010(5)