泰迪智能科技
TipDM Intelligent Technology

Python 数据挖掘建模平台

TipDM-PB

技 术 白 皮 书

地址： 广州市黄埔区科学城开泰大道 36 号 1 栋 212

网址： http://www.tipdm.com

邮箱： services@tipdm.com

邮编： 510000

电话： 020-22205718

目　　录

1 概述

1.1 背景

当今社会已经步入大数据时代，数据挖掘已经成为各应用领域的重要技术，高校数据挖掘课程的开设也应运而生，数据挖掘课程综合了多门学科知识，该课程既包括各种理论知识，又离不开相关的实践技术，整个教学过程是培养和提高学生的创新能力及综合解决问题的能力。同时，旧的教学过程理论性强，枯燥乏味，学生学习热情普遍不高，不利于学生专业能力的培养。为了改变这一情况，Python数据挖掘建模平台应运而生。

1.2 产品介绍

Python数据挖掘建模平台是由广州泰迪智能科技有限公司自主研发，面向高校数据挖掘课程教学的数据挖掘建模工具。平台使用JAVA语言开发，采用B/S结构，用户不需要下载客户端，可通过浏览器进行访问。用户可在没有Python编程基础的情况下，通过拖拽的方式进行操作，将数据输入输出、数据预处理、挖掘建模、模型评估等环节通过流程化的方式进行连接，已达到数据分析挖掘的目的。

1.3 专用术语

组件：将建模过程涉及到的输入/输出、数据探索及预处理、建模、模型评估等算法分别进行封装，每一个封装好的算法都可称之为组件。

工程：为实现某一数据挖掘目标，将各组件通过流程化的方式进行连接，整个数据流程称为一个工程。

模型：主要针对分类、回归算法而言，使用一部分数据用于训练，会得到一个模型，里面将保存算法的参数，可使用该模型对另一批数据进行验证或预测。

个人组件：用户可按照平台规定的格式编写脚本，配置相关输入、输出、算法参数，可作为平台组件，反复调用。

任务：支持定时同步数据库数据源至平台或定时运行某一工程。

2 技术特点

2.1B/S 架构

分布性：B/S架构具有分布性特点,可随时随地进行查询、浏览等业务处理。这种体系架构是在WWW和互联网技术的流行性中发展起来的,使得用户的访问不再受到平台和软件的限制,大大增加了网站系统的适用范围,满足了用户信息可见和信息共享的要求。

扩展性：B/S架构扩展简单方便,通过增加网页即可增加服务器功能。基于B/S的三层体系架构,工作人员只需使用既定的模式和方法,通过增加网页即可达到完善功能模块、提升用户体验、提高服务质量的目的。

易维护：B/S架构维护简单方便,只需要改变网页,即可实现所有用户的同步更新。基于B/S的三层体系架构比较全面地体现了网站的逻辑体系结构,在表现层与数据层之间又添加了逻辑层。正是由于逻辑层的存在,降低了网站系统对客户端和服务器端的依赖性。许多逻辑处理工作都交予中间层来完成。在后期的维护工作中,无须对三层结构中的每一层都更改,因此维护起来较简单。

共享性：B/S架构开发简单,共享性强。将逻辑处理工作交予中间层来处理,降低了开发建设工作的难度,增强了网站系统的操作性,使用浏览器进行数据的访问,降低了对访问软件的限制,加强了信息数据的共享性。

2.2Python 计算引擎

Python的定位是“优雅”、“明确”、“简单”，使用Python作为平台的计算引擎具有如下7中优势。

语法简单：Python程序简单易懂，初学者学Python，不但入门容易，而且将来深入下去，可以编写那些非常非常复杂的程序。

开发效率高：Python有非常强大的第三方库，结合Python官方库能够解决90%以上的数据挖掘问题，而不需要重头完整实现所有算法。

高度可扩展：可以与C/C++实现无缝连接，实现混合编程。

2.3 工作流

以工作流的形式展现数据挖掘流程，具有下列9大优势。

（1） 图形化、可视化设计流程图

（2） 支持各种复杂流程

（3） 组织结构级处理者指定功能

（4） B/S结构，纯浏览器应用

（5） 强大的安全性特色

（6） 表单功能强大，扩展便捷

（7） 灵活的外出、超时管理策略

（8） 处理过程可跟踪、管理

（9） 丰富的统计、查询、报表功能

2.4RESTful 接口设计

接口模块基于标准RESTful设计，用户可以方便，快捷的通过浏览器在线浏览、测试各个接口。RESTful具有如下优势。

（1） 前后端分离，减少流量

（2） 安全问题集中在接口上，由于接受json格式，防止了注入型等安全问题

（3） 前端无关化，后端只负责数据处理，前端表现方式可以是任何前端语言（android，ios,html5）

（4） 前端和后端人员更加专注于各自开发，只需接口文档便可完成前后端交互，无需过多相互了解

（5） 服务器性能优化：由于前端是静态页面，通过nginx便可获取，服务器主要压力放在了接口上

3 运行环境

为保证本平台在教学中可正常使用，以 40 个用户并发使用为标准，提供标准运行环境如表 3-1 所示。

表 3-1 软硬件运行环境说明

	名称	配置要求 （CPU、内存、硬盘）	数量	用途
硬件	服务器	CPU：不小于 2 颗 CPU，8 核 16 线程。建议采用 Xeon E5-2620 v4, 2.1GHz，8T/16C，带宽 6.4GB/s/UNIX，tpmc≧16500 内存：不小于 16G 内存。建议至少采用 2*8GB RDIMM DDR4 2400MT/s 硬盘：不小于 1T。建议采用 SAS 硬盘 1T	2	计算服务器
	服务器	CPU：不小于 2 颗 CPU，8 核 16 线程。建议采用 Xeon E5-2620 v4, 2.1GHz，8T/16C，带宽 6.4GB/s/UNIX，tpmc≧16500 内存：不小于 8G 内存。建议至少采用 2*4GB RDIMM DDR4 2400MT/s 硬盘：不小于 5T 硬盘。建议采用 SAS 硬盘 5T	1	数据服务器

软	软件名称	版本号	生产商/来源	用途

件	CentOS	6.9 以上	The CentOS Project | Legal | Privacy	服务器操作系统
	PostgreSQL	9.5 及以上	The PostgreSQL Global Development Group	数据存储
	Redis	3.0 及以上	Redis Labs	缓存
	Python	3.6 及以上	Python Software Foundation	计算引擎
	JDK	1.8 及以上	Oracle	JAVA 运行环境
	Tomcat	8.0 及以上	The Apache Software Foundation	网络服务
网络环境	千兆内网			
其它				

4 产品架构

TipDM 大数据挖掘建模平台系统架构如图 4-1 所示

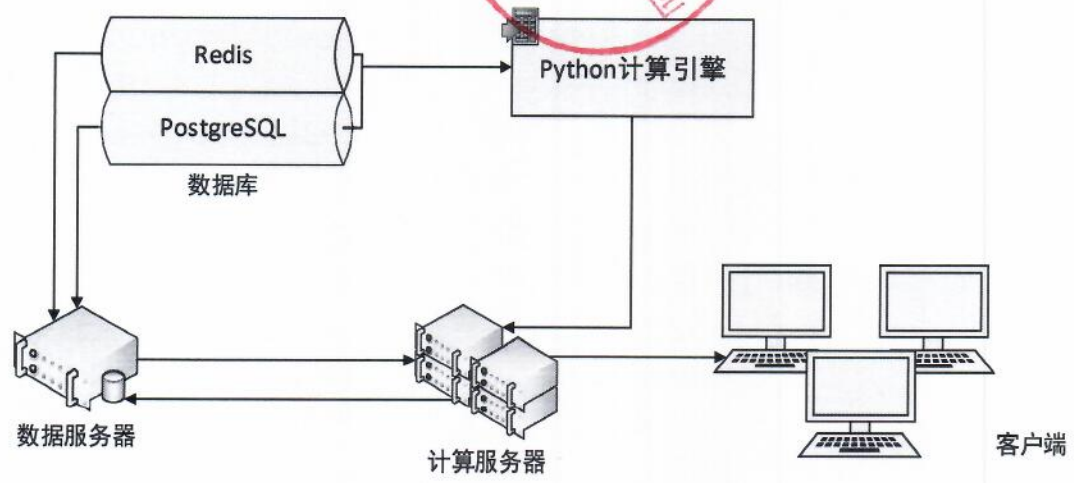

图 4-1 系统架构图

平台 Web Server 主要有两大功能：接收前端用户访问请求，对后台数据源或算法调用进行操作。

TipDM 大数据挖掘平台数据源采用 2 种数据库，分别是 Redis、PostgreSQL。 Redis 主要存储系统运行时临时消息，可以高效进行消息的传递及查询。PostgreSQL 主要存储用户数据，当用户数据较少时（这时用户一般运行的是单机算法来对其数据进行处理），使用 PostgreSQL 来存储用户临时数据，同时在工程流程中运行生成的临时表也存储在 PostgreSQL 中。

5 产品功能

Python 数据挖掘建模平台主要包括数据源管理模块、组件管理模块、工程管理模块、任务调度、模型管理及实现各个具体功能的子模块。

各模块的具体关系如表 5-1 所示。

表 5-1 产品功能模块关系

模块	功能描述
首页	查看社区、查看入门、通过模板创建工程
数据源	上传数据源、数据分享、数据预览
工程	创建工程、运行工程、参数设置
系统组件	新增系统组件、编辑系统组件、编辑源码
个人组件	新增个人组件、编辑个人组件、编辑源码
模型	导入模型、分享模型、模型预测
任务	新建工程任务、新建数据源任务