基于策略模式的中医数据挖掘平台①

张连育[1,2] 吕 立[2] (1.中国科学院 研究生院 北京 100049;
2.中国科学院 沈阳计算技术研究所 系统与软件实验室 辽宁 沈阳 110171)

摘 要: 随着数据挖掘技术的发展和中医信息化的逐渐深入，很多数据挖掘方法被应用到中医研究领域。针对面向对象软件设计模式中的策略模式在数据挖掘科研软件平台设计开发上的应用进行了研究，并提出了平台设计概要。在此基础之上，提出了一种中医数据挖掘研究的思想方法：将中医问题(数据)封装、将数据挖掘方法(算法)封装，实现统一的接口，从而实现在某一类中医问题中尝试不同的数据挖掘方法、将某一种数据挖掘方法应用于不同的中医问题。基于上述思想方法，实现了中医数据挖掘平台，用于中医相关领域的数据挖掘研究。

关键词: 数据挖掘；中医；策略模式；软件重用；中医数据挖掘平台

Traditional Chinese Medicine Data Mining Platform Based on Strategy Pattern

ZHANG Lian-Yu[1,2], LV Li[2]
(1.Graduate School of the Chinese Academy of Sciences, Beijing 100049, China; 2.Shengyang Institute of Computing Technology, Chinese Academy of Sciences, System and Software Lab, Shenyang 110171, China)

Abstract: With the development of data mining technology and traditional Chinese medicine information system, many data mining methods have been applied to traditional Chinese medicine research. This paper discusses the application of strategy pattern (one of object oriented software design patterns) on data mining scientific research software platform design and development, and the design summary of the platform. On this basis, a sort of thoughtway of traditional Chinese medicine data mining study is proposed, which is: encapsulating traditional Chinese medicine problems(data), and encapsulating data mining methods(algorithm), which realize uniform interface, thus researchers can try different data mining methods in a sort of traditional Chinese medicine problem and use a data mining method in different traditional Chinese medicine problems. Based on the thoughtway above, this paper realizes a traditional Chinese medicine data mining platform for Chinese medicine data mining research.

Keywords: data mining; traditional Chinese medicine; strategy pattern; software reuse; traditional Chinese medicine data mining platform

1 引言

设计模式是一套被反复使用、经过分类的代码设计经验的总结。使用设计模式，可以重用代码、让代码更容易地被他人理解，而且能保证代码可靠性。策略模式定义一系列的算法并封装起来，使其可独立于客户而替换、变化[1]。

① 基金项目:国家重点基础研究发展计划“九七三”项目(2006CB504800)
收稿时间:2010-03-03;收到修改稿时间:2010-03-29

数据挖掘，又称数据中的知识发现，其目的是自动或方便地提取隐藏或记录在数据中的代表知识的模式[2]。目前数据挖掘在 IT、银行、保险、医药和零售等行业以及生物、天文等科研领域都有应用，并出现了一些商用数据挖掘软件产品。

中医文化源远流长，著作浩如烟海，如何对这些文献进行高效整理、研究成为中医科研重要课题。另外，目前的中医研究中的药理研究、临床试验、动物实验，以及医院的病例诊断、治疗记录等都会产生大量的数据需要处理、研究。数据挖掘在此大有可为。

目前已有一些数据挖掘方法应用到中医研究中，但只局限于具体问题，即数据挖掘某种方法与中医研究中某一问题是一一对应的关系。虽取得一定成果，但对于探索性试验存在很大局限性。为此，本文首先探讨策略模式在数据挖掘科研软件平台设计开发上的应用，并提出平台设计概要。此基础上设计实现能够在问题和方法间实现灵活多对多关系的中医数据挖掘平台：将中医问题(数据)封装、将数据挖掘方法(算法)封装，实现：(1)某一问题采用不同方法解决，对比不同方法的解决效果；(2)某一种方法应用于不同的问题，最大限度发挥某一种方法的功效。

2　中医数据挖掘应用现状

目前医疗系统信息化已取得了很大的进展，医院信息管理系统、实验室信息管理系统、医学影像系统等已广泛应用于医院。中医现代化过程中，也建立了很多数据库，例如中医药科技信息数据库等。但以上应用大都仅限于数据存储及简单处理，缺乏数据分析。中医学具有系统性、整体性、复杂性、不确定性等特点，适合采用数据挖掘从整体观入手的研究方法。已有一些相关研究：陈明等[3]将关联规则应用于中医疾病证候诊断中。秦中广[4]等将粗糙集应用于中医类风湿证候诊断中。刘晋平[5]将数据挖掘应用于中医脉诊研究。

3　中医数据挖掘平台的理论基础和技术架构

3.1 设计模式与策略模式

面向对象软件设计模式是对被用来在特定场景下解决一般设计问题的类和相互通信的对象的描述[1]，分为创建型模式、结构型模式、行为模式等。策略模式是对象行为型模式的一种，核心意图是定义一系列的算法，把它们一一封装起来，并且使它们能够互相替换。策略模式使算法可以独立于使用它的客户而变化[1]。

3.2 数据挖掘主要方法

数据挖掘的主要方法包括以下 6 种：特征化与区分[2]、简档[6]、关联规则[2,6]、分类与预测[2,6]、聚类与离群点分析[2,6]、演变分析[2]。每种方法包含若干算法。

3.3 基于 JSP 的 MVC 架构

作为 Web 开发架构的一种，此架构将服务器端分为 3 个逻辑单元：模型(M)、视图(V)、控制器(C)。通常服务器应用分为业务逻辑、表示和请求处理。模型对应业务逻辑和数据，视图对应表示，控制器对应为请求处理[7]。视图采用 JSP 网页，模型和控制器用 Java 开发，控制器使用 Servlet 类。服务器采用 TOMCAT。

4　策略模式在数据挖掘科研平台上应用研究

数据挖掘作为计算机科学研究中的一个新兴领域，其研究存在以下两个方面：(1)数据挖掘自身理论的完善；(2)拓展数据挖掘技术的应用领域。

目前数据挖掘的理论方面已经形成了一个较为完善的体系，但在拓展数据挖掘技术的应用领域方面，由于应用行业问题形式的千差万别，很多情况下不易找到问题和方法之间的一一对应关系。因此，设计一个针对于数据挖掘应用研究的科研软件平台有其必要性：(1)从应用问题的角度而言，同一个应用问题需要借助数据挖掘理论中不同的思想方法、算法来进行科研尝试，以便观察、比较其应用效果。(2)从数据挖掘的角度而言，经典理论中同一种思想方法、算法需要尝试应用于不同的问题。

该平台应该区别于商用数据挖掘工具软件。后者目的是给出数据挖掘结果。用户并不知道系统采用何种算法，甚至不了解数据挖掘本身，他们感兴趣的只是结果。

针对于数据挖掘应用研究的科研软件平台侧重于提供给用户以下功能：(1)能够集成现有数据挖掘思想方法、算法。(2)能够导入数据挖掘应用问题，包括相应的数据。(3)能够灵活地选择问题和方法之间的组合。(4)能够比较不同的方法解决同一问题的效果，并显示给用户。(5)能够展示同一方法如何解决不同问题。

基于以上论述，策略模式非常适合数据挖掘应用

研究的科研软件平台的架构设计：

(1) 基于策略模式的思想，按数据挖掘方法/算法理论体系(如图 2)，将各种数据挖掘方法/算法封装，使它们能够互相替换，实现统一的数据处理接口，实现同一应用问题使用不同的方法/算法进行尝试。

图 2 数据挖掘方法/算法理论体系(部分)

(2) 基于策略模式的思想，在一定粒度上将各种应用问题/数据封装，实现统一的算法调用接口，实现同一数据挖掘方法/算法在不同的应用问题上进行尝试的目的。

(3) 结果显示。对于数据挖掘不同的方法/算法，其结果显示形式不相同。基于策略模式的思想，将各种显示形式封装，实现统一的接口，从而实现可视化形式上的丰富性。

根据以上分析，该平台分为以下几个功能模块：

(1) 数据挖掘模块：管理数据挖掘各种方法/算法，实现添加、删除、修改方法/算法。各方法/算法以可执行代码的形式组织，能够实现数据挖掘操作。

(2) 问题/数据管理模块：管理应用问题的数据源，实现添加、删除、修改应用问题数据源。该模块负责与数据库交互。

(3) 服务定制模块：为用户提供数据挖掘方法/算法的选择功能、应用问题的选择功能，用户选择一个“应用问题/数据、数据挖掘方法/算法”二元组后，可以通过该模块启动一次数据挖掘服务。

(4) 结果显示模块：显示数据挖掘效果。

在中医数据挖掘研究中，可以将中医问题(数据)封装、将数据挖掘方法(算法)封装，实现统一的接口，实现在某一类中医问题中尝试不同的数据挖掘方法、将某一种数据挖掘方法应用于不同的中医问题，便于中医研究者比较、分析。

5 平台需求分析及模块设计与实现

5.1 中医数据挖掘平台需求分析

此平台基于中医数据挖掘实际需要设计开发，结合第 4 章中提出的思想方法，主要需求如下：

(1) 数据录入与数据预处理。中医数据种类繁多，包括结构化和半结构化数据，不同问题对于数据考察角度也不同。为实现某一问题分别采用不同方法解决，此平台需实现数据结构化录入和逻辑模式[8]层面一致化封装。

(2) 数据挖掘方法的灵活调用。后台组件需将数据挖掘各方法封装，提供对 1 中的数据库进行操作的统一接口，实现问题(数据)与方法之间多对多调用关系。

(3) 效果展示界面。此平台需实现计算可视化，为用户展示某一数据挖掘方法解决某一中医问题的效果。因中医问题差异性大，而数据挖掘方法[2]组织条理清晰，故界面以数据挖掘主要方法组织。界面内容包括文字、图示。

5.2 中医数据挖掘平台功能模块设计与实现

为了实现数据显示、数据处理、数据库读写的分离，充分发挥 MVC 架构，功能模块(如图 3)如下：

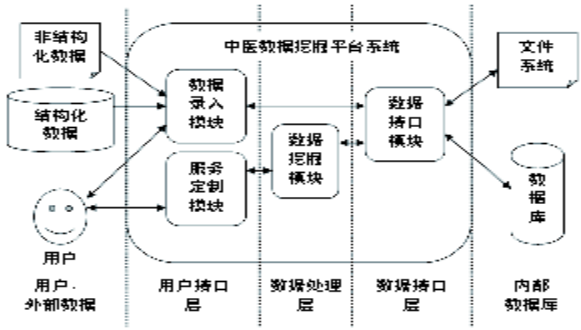

图 3 平台体系结构

(1) 数据录入模块。实现将半结构化的数据预处理为结构化的功能。工作流程：(1)用户注册一个新的数据库表，并为其注册属性，然后对属性进行规范化编码输入，以 XML 文件作为数据库缓冲文件；(2)系统将半格式化数据导入数据库，并由用户根据规范化编码为数据条目作结构化标引，存入数据库。实现对不同的数据做统一逻辑模式的封装。凡涉及数据库读写的操作均调用数据接口模块。数据类的关系见图 4：第 1 层为抽象类，最底层为接口，实现与图 5 中的 DataOpInterface 兼容。每一个数据类继承自上层抽象父类，并实现 DataInterface 接口。此模块工作在

用户接口层，需要与用户进行交互，采用 JSP 页面作为前台显示界面，Servlet 类实现后台调用。

(2) 数据挖掘模块。由一系列可注册添加的后台组件构成。这些组件针对模块 1 中封装的数据提供兼容的接口，组织形式如图 5 所示。图中第 1 层、第 2 层为抽象类。最底层为接口，实现与图 4 中的 DataInterface 兼容。工作流程：用户通过服务定制模块选定某一数据、某一数据挖掘方法，启动数据挖掘模块。该模块调用数据接口模块读取相关数据，进行处理，结果交付服务定制模块。各数据挖掘方法类继承自上层抽象父类，实现 DataOpInterface 接口。此模块工作在数据处理层，这些组件用 Java 语言开发，不直接与用户交互，被 Servlet 类后台调用。

(3) 服务定制模块。实现 2 个功能：(1)提供一个与图 4 中第 2 层、图 5 中第 2 层(以及以下各算法层)对应的用户选择功能，即一个已经注册的数据类型和已经注册的数据挖掘方法的选择列表；(2)接收数据挖掘模块得到的结果并显示给用户。由于每种数据挖掘方法处理对象的角度不同，故结果的展示形式也不同，图 5 中第 2 层的每种方法对应此模块中的一个显示组件。此模块工作在用户接口层，需要与用户进行交互，采用 JSP 页面作为前台显示界面，Servlet 类实现后台调用。

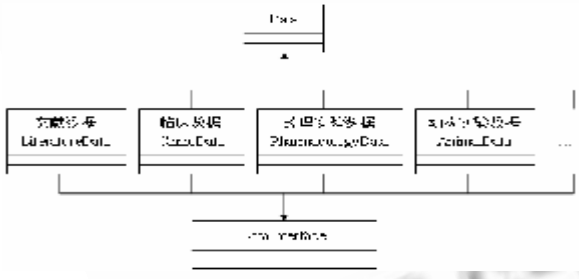

图 4　数据类关系图

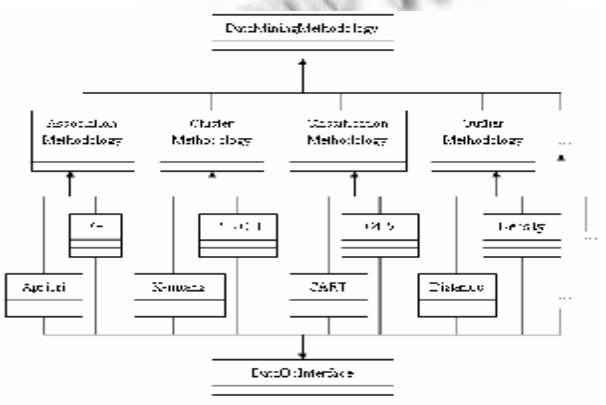

图 5　数据挖掘方法类关系图

(4) 数据接口模块。负责一切与读写数据库相关的操作，工作在数据接口层。采用 JDBC 接口，将 JSP 页、XML 文件、其他数据库写入后台数据库、从后台数据库中读取数据。

(5) 其他说明：(1)数据录入模块对可以直接应用的结构化数据提供接口，通过数据接口模块直接导入数据库。(2)为了提高数据挖掘效率、存储挖掘结果、引入中医领域已有知识等，需要在内部数据库中建立知识库，便于交互。

6　平台功能及性能评价

6.1　基于策略模式的功能实现

(1)数据挖掘是中医研究采用的新方法，具体问题需要尝试多种数据挖掘方法。例如问题 1：通过中医文献研究某病机与何症状相关。可用关联规则或分类解决。可注册一属性包括病机、症状(图 6)的数据库，导入文献条目并标引，分别定制关联规则和分类相关的方法进行数据挖掘。

图 6　症状编码

结果	置信度
舌苔薄(白)、吞酸或泛酸、舌质(淡)红、脘肋胀痛=>气机不和证	87%
恶心、呕吐=>痰湿中阻证	90%
舌苔薄(白)、脉细、苔有齿印=>气机不和证	90%
脉细、苔有齿印=>气机不和证	100%
舌质偏红、口干(欲饮)、肌肋胀痛=>脾胃气虚证	80%
胃脘嘈杂、灼热、吞酸或泛酸=>胃阴虚证、脾胃湿热证	100%
脘肋疼痛、性情急躁或欠佳=>肝胃不和证	83%
胃肌嘈杂、灼热、吞酸或泛酸=>胃阴虚证	100%
灼热、吞酸或泛酸=>脾胃湿热证	100%
便溏、脉细弦、口干(欲饮)、脘肋胀痛=>脾胃气虚证	81%
舌苔薄(白)、便溏、脉细弦、口干(欲饮)=>脾胃气虚证	100%

图 7　关联规则结果展示

(2) 中医是数据挖掘应用的一个新领域，数据挖掘的同一种方法可能会应用到不同的中医问题中。譬如关联规则除问题 1，还可应用到问题 2：通过中医治疗用药数据研究某种“证”(疾病)所用方剂的频率高低。

(3) 数据挖掘结果可视化一般包括文字、图示两部分。每种方法的显示形式不同。所以需要应用策略

模式，在统一的显示接口下采用多种显示组件。图 7 为关联规则显示。

6.2 性能评价

平台开发目的是为中医研究提供计算机辅助工具。对于数据挖掘，目前中医研究主要关注各种挖掘方法在不同中医问题中尝试得到的结果，评价其贡献性，而非具体算法性能。

目前此平台工作于配置为主频 2.5GHz 以上的双核 CPU、4GB 内存、Windows XP、Tomcat 5.0、Oracle 9i 的服务器上，可顺利完成 103 量级的数据规模的数据挖掘计算。

6.3 平台先进性评价

目前，中医数据挖掘研究大多是实验性质，即将个别数据挖掘方法应用于具体的中医研究问题，参见本文第 2 章中所介绍。也出现了一些中医数据仓库、数据挖掘系统，但仍属于本文第 4 章中提到的商用数据挖掘工具软件范畴，不能达到问题和方法间实现多对多灵活切换的科研实验效果。在这方面，应用策略模式设计的本平台相比于上述具有优势。

7 结语

本文针对策略模式对于数据挖掘应用研究的科研软件平台架构的适用性进行了论述，并提出设计概要。在此基础上实现了一个中医数据挖掘科研软件平台，其具有很强的灵活性优势：在问题端和解决方法端都可以实现开放性的添加、删除、更新，并可以通过服务定制进行任意一种方法解决任意一类问题的试验，避免问题和方法之间的互相束缚，满足科研的要求。

在今后的工作中，将继续对此平台从以下几个方面进行改进：(1)丰富知识库，使数据挖掘模块的工作效率更高；(2)升级显示界面，实现由 2D 转化成 3D；(3)增加算法组件，增强平台的灵活性。

最后，这种基于策略模式，在问题端和方法端都进行封装，从而实现问题和方法之间多对多灵活选择试验的设计思想，可以应用在很多科研平台上。

参考文献

1 Erich Gamma. Design Patterns: Elements of Reusable Object-Oriented Software. New Jersey: Addison Wesley Longman, 1995.

2 韩家炜,堪博.数据挖掘概念与技术.第 2 版,范明,孟小峰译,北京:机械工业出版社, 2007.3.

3 陈明,张书河.关联规则在中医疾病证候诊断中的应用.中华医学丛刊, 2004,4(5):14－16.

4 秦中广,毛宗源,邓兆智.粗糙集在中医类风湿证候诊断中的应用.中国生物医学工程学报, 2001,20(4):357－363.

5 刘晋平,黄宇虹,陆小左.数据挖掘在中医脉诊中的应用.天津中医药学院学报, 2003,22(3):9－10.

6 贝瑞,莱诺夫.数据挖掘技术;市场营销、销售与客户关系管理领域应用.第 2 版,别荣芳,尹静,邓六爱译,北京:机械工业出版社, 2006.7.

7 鲍格斯坦. JSP设计.第3版,林琪,朱涛江译,北京:中国电力出版社, 2004.

8 萨师煊.数据库系统概论.第 3 版,北京:高等教育出版社, 2002.