

肺部影像组学数据库系统设计与实现

袁慧敏, 龚敬, 聂生东

上海理工大学医学影像工程研究所, 上海 200093

【摘要】近年来基于肺部CT影像组学研究已成为医学影像领域中重点发展方向之一。但是目前国内外现存肺癌相关数据库存在信息欠缺,缺乏统一数据模型,无法实现数据共享等问题。因此,本文构建了肺部影像组学数据库系统。首先,根据影像组学研究所需数据标准进行需求分析及系统功能、结构以及数据库设计;然后,在B/S架构,基于MVC模式的ThinkPHP开发框架下,利用HTML5+JavaScript+CSS技术组合完成系统前端数据的显示,PHP技术完成服务器端逻辑处理,MySQL实现对影像组学数据存储与管理;最后,运行测试结果表明,本数据库可兼容多平台实现数据共享,不仅可实现DICOM格式CT影像可视化且在补充LIDC数据库中数据项的基础上使得CT影像数据与多学科临床诊疗数据形成统一的数据模型,推进了肺癌精准诊疗研究进展。

【关键词】基于肺部CT影像组学;数据库;B/S;ThinkPHP;MySQL

【中图分类号】R319;R811

【文献标志码】A

【文章编号】1005-202X(2019)11-1284-07

Design and implementation of lung radiomics database system

YUAN Huimin, GONG Jing, NIE Shengdong

Institute of Medical Imaging Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Radiomics based on lung CT has recently become one of the most important development trends in the field of medical imaging. However, there are some problems in the existing domestic and international database of lung cancer, such as lack of information, lack of unified data model, unable to realize data sharing and so on. In order to solve the above-mentioned problems, a lung radiomics database system is constructed. Firstly, the requirement analysis and the designs of system function, system structure and database are carried out according to the data standards required in radiomics research. Then, in B/S framework, based on the ThinkPHP development framework of MVC mode, HTML5+JavaScript+CSS technology is used for the display of system front-end data, PHP technology for server-side logic processing, and MySQL for the storage and management of radiomics data. Finally, the test results show that the database can be compatible with multiple platforms to achieve data sharing. On the basis of supplementing data items in LIDC database, the database can not only realize the visualization of CT image in DICOM format, but also form a unified data model between CT image data and multi-disciplinary clinical data, thus greatly promoting the research on precision diagnosis and treatment of lung cancer.

Keywords: radiomics based on lung CT; database; B/S; ThinkPHP; MySQL

前言

作为“癌症之首”,肺癌发病年龄比以往提早了5~10年^[1]。若要提高肺癌治愈率,须早诊断、早治疗^[2]。随着精准医疗时代的到来与发展,影像组学

(Radiomics)的概念自2012年Lambin等^[3]提出以来引起了影像研究者的广泛关注,基于肺部CT影像组学通过对肺癌CT影像提取大量特征,再研究这些特征与临床数据之间的深层关系,揭示肺癌的发生、发展及临床转归规律,在预测和指导肺癌的精准诊疗研究方面具有较大的潜能^[4-6]。

数据是以上研究的基础和关键,而国内现存的相关肺癌研究数据库大多采用C/S架构,且存在界面老化、操作不便、服务对象范围较狭窄、无法实现数据共享等问题^[7]。国际上应用最广泛的肺结节检测CT影像数据库-LIDC/IDRI数据库虽然促进了肺结

【收稿日期】2019-06-08

【基金项目】国家自然科学基金(60972122)

【作者简介】袁慧敏,硕士研究生,研究方向:医学影像处理及软件工程,
E-mail: 1135963194@qq.com

【通信作者】聂生东,博士,教授,研究方向:医学图像处理与分析、核磁共振图像以及信号处理分析, E-mail: nsd@163.com

节计算机辅助诊断(Computer-Aided Diagnosis, CAD)的发展,但其未建立一个完整的数据模型,造成数据存取不便,且数据的完整性和一致性无法得到保障。同时,该数据库中每个样本病例只包含一组CT影像,无法实现肺结节病灶区域生长变化的研究。最后,该数据库中的病理诊断结果不完整,无法实现对肺肿瘤良恶性预测和鉴别的研究,且缺乏临床信息,无法利用多学科综合信息实现对肺癌的精确诊断^[8-9]。文献[10]采用VC++工具对LIDC数据库中的数据做了3点改进:①增加数据检索功能;②统一数据模型;③CT图像浏览可视化。但仅有以上改进,仍不能实现数据共享也无法满足影像组学精准诊疗的需求。

为解决以上问题,本文研究了肺部影像组学数据库系统设计及开发方法,首先进行需求分析并以此为基础进行功能模块设计,然后进行系统结构和数据库设计,最后采用B/S开发架构、MVC开发模式,利用HTML+JavaScript+CSS以及PHP技术开发数据库的前端数据检索和后台数据管理的界面和功能,接着使用MySQL数据库管理系统对各类型数据统一管理,最后利用多学科数据以及肺部CT影像数据进行系统测试。最终为肺部影像研究者提供一个以数据共享为核心,功能全面,数据全面且标准的可满足肺部影像组学精准诊疗需求的数据库系统。

1 材料与方法

1.1 肺部影像组学数据库系统设计

1.1.1 功能结构设计 作为一种面向肺部影像研究者

的新型肺部影像资源及各学科数据组织与管理的平台,肺部影像组学数据库系统与其它现存肺癌相关数据库相比,具有以下特性:①以基于肺部CT影像组学为中心,组织与管理影像资源与多学科融合数据,包括病人基本信息及病理诊断数据、CT影像诊断数据、肺结节数据、基因数据、肿瘤报告数据、治疗前最后一套和治疗后第一套CT影像,为肺部疾病量化分析、精确诊断及肺部病灶区域生长变化研究提供全面的数据基础;②基于对影像研究需求和HTTP网络传输与响应时间这两种因素的权衡考虑,系统中设置的CT影像格式有3种,包括JPG、DICOM和Zip格式,研究者可根据个人研究需求向系统请求不同格式的影像数据;③在DICOM格式影像文件中,除了可调节窗宽、窗位等功能,还可以进行数字图像的其他专业处理,包括连图像平滑与锐化、连续阈值变换、Sobel变换等,可为研究者进一步分析影像提供有效参考;④为保证数据安全,防止重要数据丢失,建立回收站机制。系统总体功能设计如图1所示。总体功能包括各学科数据检索、CT影像后处理(包括JPG和DICOM格式)、留言板、系统设置、权限设置、表格导入与导出数据库、回收站管理、数据管理。在满足以上系统功能的前提下,再进一步按照父级节点功能细分各个子模块。

1.1.2 系统结构设计 在总体功能模块设计的基础上,将数据库系统分成前台数据检索与后台数据管理两大部分。其中,用户在前台通过后台链接转到后台登录界面,再将以上两部分与底层数据库连接,形成如图2所示的系统总体结构设计。

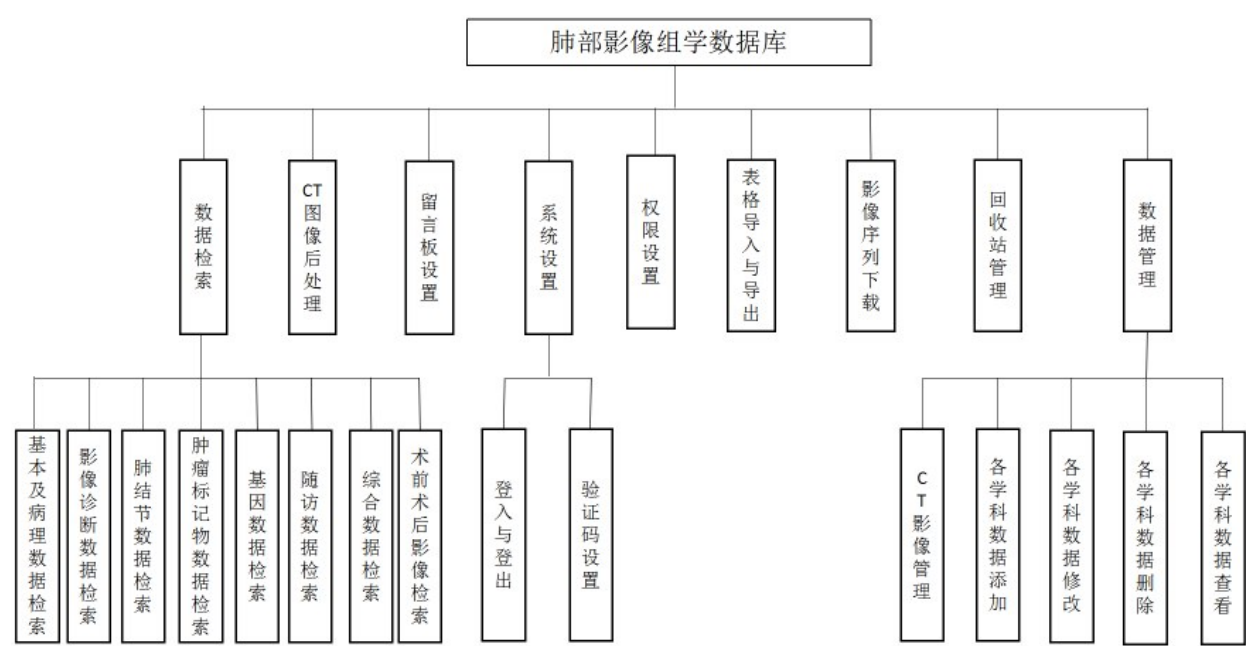


图1 肺部影像组学数据库系统功能结构设计

Fig.1 Functional structure design of lung radiomics database system

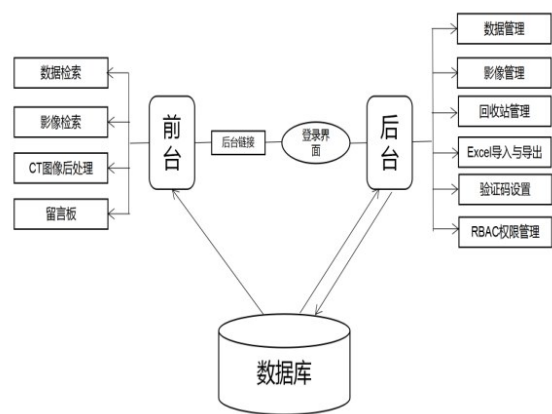


图 2 肺部影像组学数据库系统总体结构设计
Fig.2 Structure design of lung radiomics database system

前台设计包括 CT 影像(JPG 格式)的查询与后处理、影像组学相关多学科临床诊疗数据查询、留言板功能。其中,影像查询支持治疗前后两套影像同时查询;通过引入插件实现影像后处理操作,包括选定感兴趣区域、感兴趣区域上下平移、放大缩小、旋转、锁定等操作。经过后处理操作形成的 CT 影像并不覆盖原图像,以保证影像资源的可复用性。多学科数据查询包括基本及病理诊断、肺结节位置、影像诊断、基因、肿瘤数据、随访数据查询。在数据查询方式的设置中,除了包含普通查询方式,如:按病人 ID 号、性别、年龄、病理检查日期等;还包括肺结节直径大小、病理类型、TNM 分期等专业查询方式。此外,系统还支持“各学科综合数据-肺结节数据-CT 影像”递进式查询,此查询方式设计不仅能满足研究者数据需求,也可有效提高研究者的工作效率。最后,前台还设置了“留言板”功能,有利于激发用户的参与积极性,提高个性化服务水平。

后台主要负责系统所有相关数据管理与维护,其安全性对于系统的正常运行至关重要。因此系统添加了基于角色的权限访问控制(Role-Based Access Control, RBAC)管理模型设置^[11],由管理员为后台添加不同角色的用户,并给用户分配唯一的用户名、密码以及相应权限。在设计此部分功能时,共添加了 3 种角色,分别为管理员、网站编辑、普通用户。其中,管理员具有所有操作权限,包括数据的管理与维护、用户管理、角色管理、权限分配等;网站编辑的权限包括验证码字母个数与大小的设置等;普通用户只能查看各个学科和 CT 影像数据。用户需在登录界面输入正确的用户名、密码以及验证码,待系统验证后方能进入后台。

后台的数据管理主要包括各学科诊疗数据管理、CT 影像管理、Excel 数据导入与导出。其中各学

科诊疗数据管理包括数据添加、删除、修改、查看,同时为了保证数据的安全,防止用户操作失误而造成重要数据丢失,后台还添加了回收站机制,被点击“删除”的数据将会暂时进入回收站,在回收站中可将数据“还原”,或从数据库中彻底删除。在 CT 影像管理中,设置了 JPG 格式影像的上传、压缩、将上传影像自动保存至服务器以及将影像相对路径等信息保存至数据库等。此外,还设置了 DICOM 格式 CT 影像的上传及后处理操作,研究者可对每一帧 DICOM 格式的影像进行窗宽和窗位调节、利用窗口中的工具勾选感兴趣区域、对图像进行阈值变换、Sobel 变换等专业数字医学图像变换,辅助研究者进行影像分析。另外,系统还提供被压缩成 JPG 格式的 DICOM 影像序列上传与下载,以上设计思想既能满足研究者对医学影像精度的需求,又可节省系统所占空间,进而提高系统性能。最后,后台设置了 Excel 数据导入与导出功能,可一次性将表格中的数据批量添加到数据库,也可将数据库中的数据一键导入到表格中供研究者进行数据统计学分析,减轻管理员数据录入的工作负担。

1.1.3 数据库设计 数据库设计的目的不仅在于实现肺部影像组学相关数据存储与管理,使得原本松散独立的数据形成一个完整的数据模型,还包括概念设计和逻辑设计。

概念设计阶段主要分析影像组学数据库系统中存在的实体,再抽象出各实体的属性以及各实体间的关系,最终形成 E-R 模型^[12]。肺部影像组学数据库系统中各项数据形成的 E-R 模型如图 3 所示,共包括 8 种实体,分别是病人、基本信息记录及病理诊断报告、CT 影像诊断报告、CT 影像序列、基因诊断报告、肿瘤标记物检查报告、肺结节、术后随访记录。病人与基本及病理数据之间的关系是 1 对 1;病人与 CT 影像、基因以及肿瘤标记物检查报告之间的关系也是 1 对 1;病人与 CT 影像之间的关系是 1 对多;CT 影像序列与影像报告之间的关系为多对 1;影像报告与肺结节之间关系为 1 对多;肺结节与随访数据之间的关系为多对 1。

逻辑设计是将已经设计好的 E-R 模型转换为关系模型,其能够更加清晰地描述系统业务与需求之间的关系^[13]。因此本数据库包括基本及病理数据表、肺结节数据表、影像诊断表、基因报告表、肿瘤报告表、CT 影像路径表(包括治疗前后)、随访数据表,以上各表及其包含字段,如表 1 所示(治疗前后的 CT 影像路径表设置相同,故二者只做一次说明)。其中基本及病理诊断表为主表,其字段“序号”设置为主

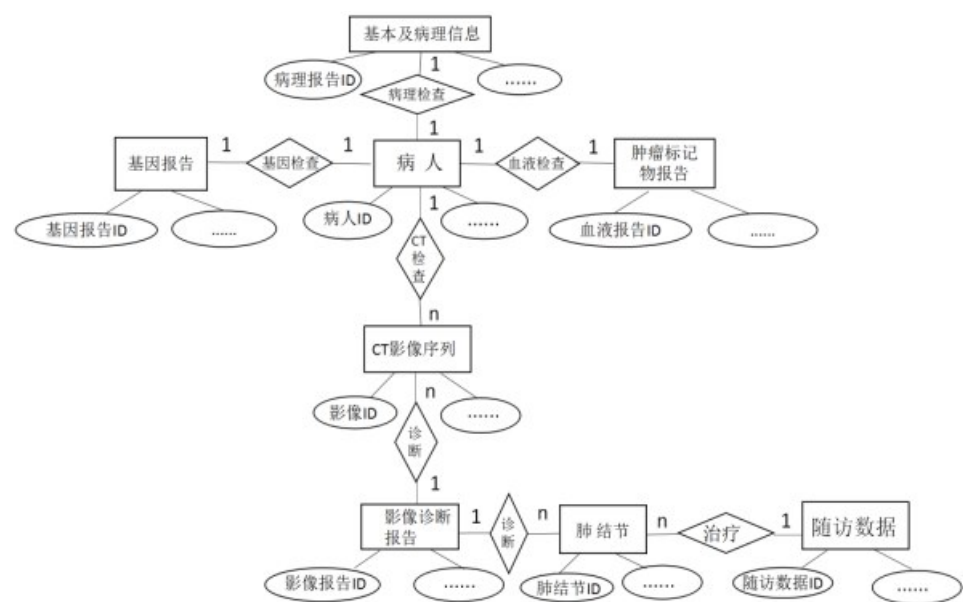


图3 各项数据类E-R模型

Fig.3 E-R model of various data classes

表1 肺部影像组学数据库逻辑设计

Tab.1 Logical designs of lung radiomics databases

数据表名称	字段1及其数据类型	字段2及其数据类型	字段3及其数据类型	字段4及其数据类型	其余字段
基本及病理诊断表	序号 PK, int (6)	年龄 tinyint (2)	吸烟史 varchar (2)	病理诊断 text
肺结节数据表	序号 PK, int (6)	肺结节直径 int (2)	良/恶性 varchar (2)	所属病人序号 FK, int (6)
CT影像诊断表	序号PK, FK int (6)	CT影像时间 int (10)	影像诊断报告 text	手术切除部位 text
基因报告表	序号PK, FK int (6)	EGFR varchar (5)	ALK varchar (5)	KRAS varchar (5)
肿瘤报告表	序号PK, FK int (6)	SCC decimal (4,3)	CEA decimal (4,3)	CA15-3 decimal (4,3)
CT影像路径表	序号PK, int (6)	缩略图路径 char (200)	创建时间 varchar(50)	所属病人序号 FK, int (6)
随访数据表	序号PK, FK int (6)	存活期 tiny (5)	死亡状态 varchar (5)	备注 tinytext

键,其余各表都有某字段设置为外键,设置为外键字段的数值只能来源于主表中的主键值,由此,各数据表形成了一个相互联系的完整数据模型。本数据库逻辑设计符合3NF设计要求。

1.2 实现技术

1.2.1 总体开发方案 本系统采用B/S开发架构,并选用基于MVC模式的框架ThinkPHP完成系统开发。肺部影像组学数据库系统架构如图4所示。

MVC模式适用于Web开发,它将系统分成互不影响的Model模型(M)层、View视图(V)层、Controller控制器(C)层^[14]。以上3层互不影响,有利于提高开发程序的可复用性。但由于该模式本身的复杂性,若直接基于该模式开发会造成开发难度增加、周期延长等问题。本系统采用基于MVC模式的开发框架ThinkPHP进行系统的开发,其支持简洁轻便的ORM操作,配合简便的CURD操作;且具有支

持视图模型和关联模型、可轻松实现多表关联与查询。

1.2.2 关键技术 在基于MVC的设计模式下,需对M层、V层、C层分别进行数据库系统的开发。首先,V层面向用户,采用HTML、CSS、JavaScript技术进行前端界面的设计与开发。C层负责用户请求的逻辑处理,采用PHP技术,完成对研究者所发请求的处理,并连接数据库,将肺部影像组学数据库中的数据返还给研究者。PHP是一种运行在Web服务器端的脚本语言,目前采用该技术开发的Web项目约占82.3%^[15]。除了具有免费开源、支持跨平台运行等优势,其还可将用户经常访问的PHP程序驻留在内存中,因此再次访问该程序时无需重新编译,可显著提高系统的运行效率与性能。M层采用MySQL技术^[16],MySQL是一种关系型数据管理系统,不仅具有可移植性,支持多线程模式,且由于其源代码开放,对PHP有很好的支持,经常作为Web开发者的首选数据库技术。

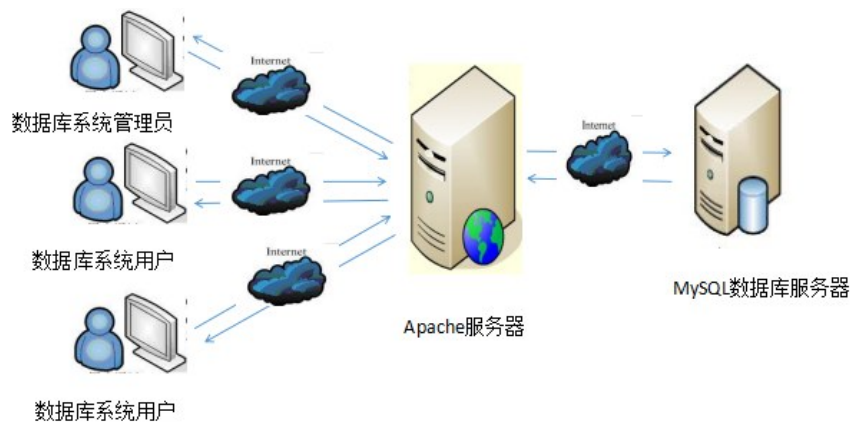


图4 肺部影像组学数据库系统开发架构

Fig.4 Development framework of lung radiomics database system

2 测试结果

本系统已发布到实验室内 Apache Web 服务器上^[17-18],使用 Linux、Windows、Mac OS 等不同平台对

本系统进行测试,其中,图5为系统首页面,图6为后台登录页面,图7为 Excel 下载页面,图8为 CT 图像 Sobel 滤波处理页面。



图5 系统首页面

Fig.5 System homepage

测试结果表明,本系统可在 HTTP 协议^[19]下顺利实现跨平台运行,并可有效保证数据安全性和一致性。

3 小结

肺癌是我国“第一癌”,其发病率在所有恶性肿瘤中位居首位。因而利用影像组学相关方法对肺癌进行精准诊疗已经成为医学影像领域的研究重点。

而数据是以上研究的关键,本研究所设计的肺部影像组学数据库采用 B/S 开发架构与国内采用 C/S 架构的肺癌相关数据库相比^[20],弥补了国内相关数据库操作复杂,服务对象受限等不足。实现数据查询与管理一体化,用户无需下载任何客户端软件,只需通过浏览器即可访问,操作简便灵活、可兼容多平台。



图 6 后台登录页面
Fig.6 Login page



图 7 Excel 下载页面
Fig.7 Excel download page

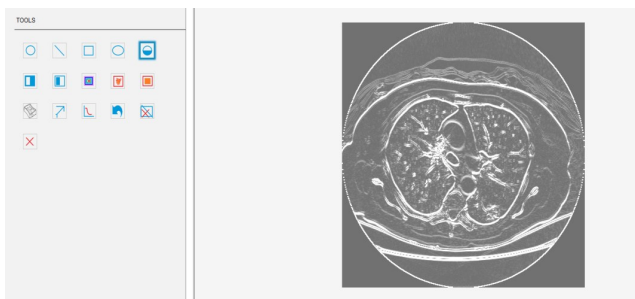


图 8 CT 图像 Sobel 滤波处理
Fig.8 Sobel filtering in CT image

【参考文献】

[1] CHEN W, ZHENG R, BAADE P D, et al. Cancer statistics in China, 2015[J]. Cancer J Clin, 2016, 66(2): 115-132.

[2] ABDOLLAHI B, SOLIMAN A, CIVELEK A C, et al. A novel 3D joint MGRF framework for precise lung segmentation [M]//Machine Learning in Medical Imaging. Springer Berlin Heidelberg, 2012: 86-93.

[3] LAMBIN P, RIOS-VELAZQUEZ E, LEIJENAAR R, et al. Radiomics: extracting more information from medical images using advanced feature analysis[J]. Eur J Cancer, 2012, 48(4): 441-446.

[4] LEE G, LEE H Y, PARK H, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art[J]. Eur J Radiol, 2016, 86: 297-307.

此外, LIDC 数据库中除影像外的数据使用文本格式和 Excel 格式保存, 不同的格式不仅造成用户下载不便, 数据的安全性和一致性也无法得到保障。故本系统采用 MySQL 数据库技术, 使治疗前后两套 CT 影像与各学科临床诊疗数据形成一个具有完整性和一致性的数据模型。最后, 本系统可向研究者提供肺癌病例治疗前后两套 CT 影像和完整的病理诊断结果及各学科临床诊疗数据, 补充了 LIDC 所含数据项的不足, 可实现对肺结节或其他肺部疾病征象及其发生、转化过程的研究及对肺癌良恶性预测和鉴别的研究, 更重要的是可较大程度上推进基于肺部 CT 影像组学对肺癌的精准诊疗研究。

[5] PAREKH V, JACOBS M A. Radiomics: a new application from established techniques[J]. Expert Rev Precis Med Drug Dev, 2016, 1(2): 207-226.

[6] YIP S S, AERTS H J. Applications and limitations of radiomics[J]. Phys Med Biol, 2016, 61(13): R150-R166.

[7] 陈嘉宝, 吴兆红. 肺癌数据库系统的研究进展[J]. 国际医药卫生导报, 2011, 17(18): 2220-2223.

CHEN J B, WU Z H. Research progress of lung cancer database system[J]. International Medicine and Health Guidance News, 2011, 17(18): 2220-2223.

[8] The Lung Image Database Consortium. NBIA at CBIIT image collections [EB/OL]. (2009-12-01) [2009-12-06]. <http://wiki.gov/display/CIP/LIDC>.

[9] National Cancer Institute. Lung image database consortium (LIDC) [EB/OL]. <http://iamging.cancer.gov/programsandresources/InformationSystems/LIDC>.

[10] 林红利, 陈真诚, 易三莉, 等. 肺癌图像数据库及可视化工具的建立[J]. 生物医学工程学杂志, 2011, 28(6): 1080-1084.

LIN H L, CHEN Z C, YI S L, et al. Development of a lung cancer image database and visualization toolkit[J]. Journal of Biomedical Engineering, 2011, 28(6): 1080-1084.

[11] BARKHA P, SAHANI G. Flexible attribute enriched role based access control model[C]//2017 International Conference on Information, Communication, Instrumentation and Control, 2017: 1-6.

[12] GETTA J R, PAN Z. Performance-based quality evaluation of database conceptual schemas [C]//2017 4th International Conference on

- Information Science and Control Engineering. IEEE, 2017: 538-544.
- [13] LIYANAGE N H. Advanced query model design concept to support multi-dimensional data analytics for relational database management systems[C]//2017 International Conference on Big Data Analytics and Computational Intelligence, 2017: 432-435.
- [14] ZHANG S, LIU Z. Research on the construction and robustness testing of SaaS cloud computing data centerbased on the MVC design pattern [C]//2017 International Conference on Inventive Systems and Control, 2017: 1-4.
- [15] GOPE D, SCHLAIS D J, LIPASTI M H. Architectural support for server-side PHP processing [C]//IEEE 44th Annual International Symposium on Computer Architecture. IEEE, 2017, 1: 507-520.
- [16] ATCHARIYACHANVANICH K, NALINTIPPAYAWONG S, PERMPOOL T, et al. Development of a MySQL sandbox for processing SQL statements: case of DML and DDL statements[C]// International Joint Conference on Computer Science and Software Engineering, 2017: 1-6.
- [17] MOCKUS A, FIELDING R T, HERBSLEB J D. Two case studies of open source software development: Apache and Mozilla[J]. Acm Trans Softw Eng Methodol, 2002, 11(3): 309-346.
- [18] RUSSELL J. Getting started with impala: interactive SQL for Apache Hadoop[M]. Sebastopol, CA: O'Reilly Media, 2018.
- [19] ZHAO S, LI Z, MEDHI D, et al. Study of user Qoe improvement for dynamic adaptive streaming over HTTP (MPEG-DASH)[C]// 2017 International Conference on Computing, Networking and Communications. IEEE, 2017.
- [20] PIRAHESH H, LEUNG T Y, HASAN W. A rule engine for query transformation in starburst and IBM DB2 C/S DBMS[C]//Proceedings 13th International Conference on Data Engineering. IEEE, 1997.

(编辑:薛泽玲)