第 37 卷 第 1 期
Vol.37 No.1

计 算 机 工 程
Computer Engineering

2011 年 1 月
January 2011

• 软件技术与数据库 •　　文章编号：1000—3428(2011)01—0063—02　　文献标识码：A　　中图分类号：TP311.13

肺癌数据库的改进及其集成工具的设计实现

王伟胜，林红利

(湖南大学计算机与通信学院，长沙 410082)

摘　要：针对肺癌计算机辅助诊断研究中使用的 LIDC 数据库缺乏统一的数据模型、不能提供数据的有效性检查以保证数据的一致性和完整性等问题，使用 XML 技术对数据模型进行改进。在此基础上提出数据库集成工具的设计方案，包括关键类及关键功能的实现思路。实践证明，该工具可以方便地完成结节的显示、检索以及对结节检测算法有效性的比较，提高研究人员的工作效率。

关键词：计算机辅助诊断；肺癌图像数据库；XML 语言；集成工具

Improvement of Lung Cancer Database and Design and Implementation of Its Integrated Tool

WANG Wei-sheng, LIN Hong-li

(School of Computer and Communication, Hunan University, Changsha 410082, China)

【Abstract】According to the problems in the use of Lung Image Database Consortium(LIDC) database in lung cancer computer aided diagnosis research, such as the lack of unified data model and the mechanism of data validity to warrant the consistency and union of data, this paper proposes an improved data model with XML technology. The design scheme of an integrated tool including main classes and the realization of its key problems are explained and analyzed. Practice shows that the tool has the ability to visualize and retrieve lung nodule as well as to evaluate the nodule detection algorithms, and it can improve work efficiency of researchers.

【Key words】computer aided diagnosis; lung cancer image database; XML; integrated tool

DOI: 10.3969/j.issn.1000-3428.2011.01.022

1　概述

肺癌是当今对人类健康与生命危害最大的恶性肿瘤之一。CT 扫描已成为早期肺癌筛查和诊断最重要的手段，但是 CT 扫描产生的大量 CT 图像直接导致了医生工作量及误诊和漏诊几率的增加。基于医学影像的计算机辅助诊断利用先进的计算机软硬件分析 CT 图像以发现并检测出病变特征，可以大大减少医生的工作量，有效地帮助医生对潜在的肺癌进行早期诊断，提高诊断准确率，减少漏诊的发生，因此，对提高肺癌的诊断水平有重要的意义[1]。

为了提高肺癌计算机辅助诊断的研究水平、为不同肺癌计算机辅助诊断研究小组的研究算法提供可比性和可重复性，美国国家癌症研究会(NCI)陆续发布了一个可以通过互联网免费下载的肺部 CT 图像数据库 LIDC(Lung Image Database Consortium)。该资源可以用来开发、训练和评价多层螺旋 CT 的肺癌计算机辅助诊断算法，同时也为不同计算机辅助诊断算法提供了标准图像[2]。目前世界上很多研究机构都是使用该数据库进行肺癌计算机辅助诊断算法的研究。

LIDC 数据库发布时间不长，还没有支持该数据库的集成工具；而且该数据库中作为评估计算机辅助诊断算法性能“金标准”的专家诊断结果是采用 XML 文件的形式提供的，为研究人员快速、直观地评价研究算法的有效性带来了很大的不便。因此，开发一个能方便研究人员对 LIDC 数据库中的病例图像进行管理、检索以及可视化的集成工具很有必要。

2　改进的 LIDC 数据模型

2.1　LIDC 数据库

LIDC 是美国 NCI 建立的一个肺部 CT 图像数据库，用以开发、训练和评价利用螺旋 CT 进行肺癌检测和诊断的计算机辅助诊断的方法。在该数据库的基础上，不同的研究者可以对比肺癌计算机辅助检测和诊断性能及其临床诊断价值，目前它的使用越来越广泛。当前 LIDC 数据库包括 399 个病例的 30 多万张全肺 CT 扫描图像(扫描层厚 1.25 mm~3 mm，512 像素×512 像素)。每个病例组织成一个文件夹，包括 100 张~200 张 DICOM 格式的全肺 CT 扫描图像和一个作为“金标准”的专家诊断结果的 XML 格式标注文件。在标注文件中按照一定的结构给出了放射科专家对每张 CT 中出现的肺结节的定义，包括结节的主要 CT 特征、结节的位置等。

多数肺癌是以肺结节的形式表现的，肺结节通常指直径小于 3 cm 的类圆形病灶，肺结节的检测是肺癌计算机辅助诊断的核心问题之一。CT 图像可以看成是将肺按照一定的层厚切成了很多切片，每个切片形成一张 CT 图像，切片的位置由图像 z 轴坐标的位置来表示。在标注文件中以肺结节为中心记录了专家对肺结节的诊断结果。专家将结节分为直径大于 3 mm 的结节、直径小于 3 mm 的结节以及非结节 3 类，每个结节和非结节有一个唯一的编号，然后定义了结节所在的 CT 图像文件以及在图像上的具体坐标。其中，直径大于 3 mm 的结节详细记录了结节的病变特征、结节所在 CT 图像的 z 轴坐标、对应的图像文件的名字、结节在每张 CT 图像上的具体坐标位置。一个结节可能在多张 CT 图像上，文件以感兴趣区(ROI)为单位记录了在每张 CT 图像上的具体位置，一个结节可以有多个 ROI，一张 CT 图像上一个结节则只有一

基金项目：湖南省自然科学基金资助项目(07JJ6133)

作者简介：王伟胜(1972－)，男，讲师、硕士，主研方向：计算机辅助诊断，软件工程；林红利，博士研究生

收稿日期：2010-06-04　**E-mail：**jt_lhl@hnu.cn

个 ROI，一个 ROI 由组成其边界的多个顶点表示，每一个顶点用由 xCoord 和 yCoord 对组成的坐标对组成。直径小于 3 mm 的结节相比直径大于 3 mm 的结节只记录了结节的中心坐标；非结节的格式与直径小于 3 mm 的结节类似，不同的是，非结节定义以标记 nonNodulek 开始，以 locus 标记代替了 edgemap，而且 locus 只有一对。结节病变特征以 characteristics 标记开始，分别从精细度(subtlety)、内部结构(internalStructure)、钙化程度(calcification)、球形度(sphericity)、边缘(margin)、分叶(lobulation)、毛刺征(spiculation)、纹理(texture)、恶性程度(malignancy)9 个方面描述结节的病变特征，每个直径大于 3 mm 的结节由一对 characteristics 标记组成。其中，直径大于 3 mm 的结节为：

```
<unblindedReadNodule>
<noduleID>4026</noduleID>
<characteristics>
<subtlety>2</subtlety>
<internalStructure>1</internalStructure>
<calcification>6</calcification>
<sphericity>3</sphericity>
<margin>2</margin>
<lobulation>5</lobulation>
<spiculation>5</spiculation>
<texture>1</texture>
<malignancy>1</malignancy>
</characteristics>
<roi>
<imageZposition>-54.639999</imageZposition>
<imageSOP_UID>1.3.6.1.4.1.9328.50.3</imageSOP_UID>
<inclusion>TRUE</inclusion>
<edgeMap>
<xCoord>92</xCoord>
<yCoord>292</yCoord>
</edgeMap>
<edgeMap>
<xCoord>93</xCoord>
<yCoord>291</yCoord>
</edgeMap>
</roi>
</unblindedReadNodule>
```

2.2 对 LIDC 数据模型的改进

原始 LIDC 数据库以文件形式组织数据，浏览查询数据不方便，也不便于管理；而且 XML 格式的标注文件没有提供相应的表示文件格式和文档有效性检查的 XML Schema 文件。本文针对以上问题对原 LIDC 数据库进行了如下改进：

(1)建立基于 XML 的数据模型

用 XML 的形式描述原有的独立文件之间的数据模型。原 LIDC 数据库的数据可以看成“病例->DICOM 文件＋XML 文件”的层次关系，一个数据库由多个病例组成，一个病例由多个 DICOM 图像文件和一个 XML 标注文件组成，用 XML Schema 定义了 XML 表示的数据模型[3]。

(2)建立 XML 标注文件的 XML Schema

为了能对 XML 标注文件的有效性进行检查，保证数据之间的一致性和完整性，本文在设计时为 XML 标注文件建立了对应的 XML Schema。其中，直径大于 3 mm 的结节的部分 XML Schema 定义为：

```
<xsd:element name=" unblindedReadNodule ">
<xsd:complexType>
<xsd:sequence>
<xsd:element ref="noduleID"/>
<xsd:element ref="characteristics" minOccurs="1"/>
<xsd:element ref="roi" maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
</xsd:element>
<xsd:element name="roi">
<xsd:complexType><xsd:sequence>
<xsd:element ref="imageZposition"/>
<xsd:element ref="imageSOP_UID"/>
<xsd:element ref="inclusion"/>
<xsd:element ref="edgeMap" maxOccurs="unbounded"/>
</xsd:sequence>
</xsd:complexType>
</xsd:element>
```

3 集成工具功能设计

3.1 功能模块设计

集成工具主要由病例管理、结节管理、图像处理功能模块组成。其中，病例管理主要完成病例的入库、出库、查询功能；结节管理主要完成结节的检索、显示功能；图像处理主要完成 DICOM 文件的信息解析、DICOM 图像数据转换、DIB 数据生成、DIB 数据修改以及 DIB 图像显示等功能。

3.2 关键类设计

集成工具设计采用面向对象设计，其关键类包括：

(1)XML 信息提取器，主要完成从原始 LIDC 数据库提取 XML 信息组成存储 XML 文件。

(2)XML 验证器，完成对 XML 信息提取器以及 LIDC 原始数据库中 XML 标注文件的格式和内容的验证，以保证 XML 文件中数据的一致性和完整性。

(3)XML 查询器，主要功能是接收用户输入的查询请求，并对该查询请求解析生成较优的 XQuery 查询表达式，通过数据访问接口执行查询后得到返回结果[4]。

(4)DICOM 图像处理器，主要是从 DICOM 文件中提取图像数据，转换为通用计算机能显示的 DIB 图像数据，完成对图像的各种操作以及显示功能。

(5)结节提取器，主要用于接收用户输入的要提取的结节编号，生成查询请求后调用 XML 查询器获得结节的坐标信息和病变特征，最后在图像上显示。

4 关键功能实现

4.1 DICOM 图像显示

DICOM 是美国放射学会和美国电器制造商协会组织指定的用于医学图像存储和传输的标准[3]。在 DICOM 标准中采用规定的格式对各种医学信息和图像数据进行编码，形成标准的 DICOM 格式文件。其数据结构采用数据元素的存储格式，每个数据元素均由标签(Tag)、值的类型(VR)、值域的长度、值域 4 个部分组成，如图 1 所示。其中，“标签”作为数据元素的标识符唯一地定义数据元素的物理意义，如病人姓名、年龄、设备、图像数据；值的类型描述了值域的数据类型，例如字符串、浮点数、整数；值域的长度描述了值域的字节数；值域则包含了数据元素的值。

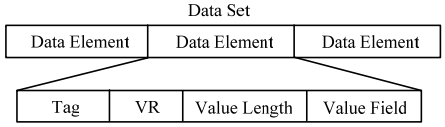

图 1 DICOM 数据结构

(下转第 68 页)

CodePlugin 采用了树状结构，树的各个结点对应代码中的各个部分。按照代码生成的顺序，为树中的每个结点设置一个 priority 属性。CodePlugin 的运行过程采用了树的前序遍历递归算法。

5 实验与分析

本文采用 DaCapo 基准测试套件 DaCapo benchmark version 2006-10 作为被监控程序，检验 FM-TSPM 工具生成的监控代码的运行时负载，从而证明工具的有效性。在实验中，对 DaCapo 的各项目使用了 HasNext、SafeEnum、HashMap 和 ClosedReader 等监控约束，这些约束来源于以往的一些运行时监控工具。

带监控能力的目标系统和初始目标系统的运行时开销的比较如表 1 所示，限于篇幅，只选取了部分数据罗列。其中，*A* 表示 Java FM-TSPM Tool 生成的监控代码与初始系统的开销比；*B* 表示手工书写的优化后的 AspectJ 监控代码与初始系统的开销比。

表 1 系统运行时的开销对比 (%)

程序	SafeEnum		HashMap		HasNext		ClosedReader	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
antlr	0.0	1.5	0.0	1.1	0.4	0.0	5.8	0.0
chart	0.0	0.0	3.6	4.8	0.0	0.0	0.0	0.0
eclipse	4.1	0.8	3.7	0.5	3.8	1.5	2.2	2.4
fop	1.2	0.6	0.0	0.0	0.8	1.5	0.0	0.0
luindex	1.6	0.2	1.2	1.8	0.3	0.0	1.7	1.1
pmd	0.0	0.0	0.0	0.0	25.4	13.7	0.0	0.0

可以看出，工具生成的 AspectJ 监控代码的运行时开销在大部分时候接近手工优化后的 AspectJ 代码，有时甚至超越后者，这充分说明工具产生的监控代码具有良好的性能。

6 结束语

本文提出了一种基于形式化监控的可信软件构造模型 FM-TSPM，用面向方面的方法将形式化逻辑编织到目标代码中，实现代码监控，为可信软件的构造和开发提出了一种全新的思路。FM-TSPM 模型目前还处在模型搭建和研究阶段，形式化解析算法和工具界面都尚待完善。目前只能解析时序逻辑语言和确定有限自动机 2 种形式语言，且在解析算法上还有很大的突破空间。另外，本模型的工具还不够完善，不能为开发人员提供一套全自动的形式逻辑导入和监控代码自动织入的功能，这将是后续研究的重点。

参考文献

[1] 林惠民, 张文辉. 模型检测理论、方法与应用[J]. 电子学报, 2002, 30(s1): 1907-1912.

[2] 唐 姗, 彭 鑫, 赵文耘, 等. 基于线性时序逻辑的动态软件体系结构模型验证[EB/OL]. (2010-06-19). http://www.se.fudan.edu.cn/paper/ourpapers/155.pdf.

[3] 郭 建, 边明明, 韩俊岗. LTL 公式到自动机的转换[J]. 计算机科学, 2008, 35(7): 241-244.

[4] 李仁点. 基于监控的可信软件构造技术的研究与实现[D]. 长沙: 国防科学技术大学, 2007.

[5] Chen Feng, Rosu G. Towards Monitoring Oriented Programming: A Paradigm Combining Specification and Implementation[C]// Proc. of RV'03. Boulder, USA: [s. n.], 2003.

编辑 顾姣健

(上接第 64 页)

LIDC 提供的 CT 图像是 DICOM 格式的，由于 DICOM 图像中图像的灰度值范围非常大(一般为 2 048 个灰度级)，通用的计算机不能支持 DICOM 图像的显示，因此需将 DICOM 格式的图像数据转换为通用计算机支持的图像格式才能显示，本文采用的是 DIB 格式。构造 DIB 文件结构的过程如下：

(1)从 DICOM 文件中读取原始图像数据，变换为能直接在显示器上显示的 0~255 的灰度数据。

(2)调整每行图像数据的字节数。判断转换后每行图像数据的字节数是否为 4 的倍数，如果不是，则需要在每行末尾补 *n* 个 0，生成新的数据区，其中，*n*=图像的列数%4。

(3)生成 DIB 数据后显示图像，然后提取当前图像上结节的 ROI 边界，用多边形表示后显示。

4.2 基于 XML 格式的结节信息检索

集成工具必须能用各种形式可视化结节和非结节的信息：大于 3 mm 的结节用红色显示该结节的形状；小于 3 mm 的结节显示为以结节为中心的边长为 5 个像素点的绿色矩形；非结节点则用以非结节为中心的边长为 5 个像素点的蓝色矩形显示。本文通过读取 DICOM 文件中表示像素间距离空间的数据元素发现：像素的距离空间在 0.55 mm~0.57 mm 之间，3 mm 的结节基本上需要用 5 个左右的像素点来表示，因此，非结节的显示使用边长为 5 个像素点的正方形来表示。

以结节为检索条件，得到分布在多张 CT 图像上的各 ROI 所在的图像文件名以及边界后，分别在对应的图像上显示；直径小于 3 mm 的结节和非结节则在得到结节的中心位置后构造一个以该位置为中心的矩形后显示。其中，每个结节 ROI 所在的 CT 图像文件的名字由 XML 标注文件中的 <imageSOP_UID>表示。

在开发中，本文使用了.NET FrameWork3.5 提供的 XPathDocument、XPathNavigator、XPathNodeIterator 等关键类实现了从 XML 文档中查询相关内容。

5 结束语

增加了 XML Scheme 后，XML 文档数据的有效性得到了保证，因此，本文开发的肺癌数据库管理的集成工具可以方便地完成结节的显示、检索以及对结节检测算法有效性的比较。实践证明，该工具可以方便地进行肺癌数据库的管理，为研究人员提供可视化的结节信息，大大提高了研究人员的工作效率，有较高的使用价值。

参考文献

[1] 魏 颖, 郭 薇, 孙月芳, 等. 面向肺癌 CAD 系统的感兴趣区域特征选择与分类算法[J]. 信息与控制, 2008, 37(4): 446-458.

[2] Al-matoll S G, McLennan G, McNitt M F, et al. Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community[J]. Radiology, 2004, 232(3): 739-748.

[3] 巢弘坤, 陈阅中. XML 数据类型验证算法的改进[J]. 计算机工程, 2009, 35(19): 53-55.

[4] 李 波, 杨卫东. XML 流上的关键字查询算法[J]. 计算机工程, 2009, 35(4): 35-37.

编辑 张 帆