第 34 卷 第 1 期 计 算 机 工 程 2008 年 1 月
Vol.34 No.1 Computer Engineering January 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)01—0259—03 文献标识码：A 中图分类号：TP393

医学影像诊断资源平台关键技术的研究

朱歆华[1]，赵大哲[1]，于亚新[2]，刘积仁[1]
(1. 东北大学软件中心，沈阳 110179；2. 东北大学信息科学与工程学院，沈阳 110003)

摘 要：研究了医学影像诊断资源平台中的一些关键技术，利用文本处理技术提取文本特征和影像处理技术提取灰度、纹理和形状特征，应用文本和影像处理相结合的技术确定影像中病灶特征，使用索引技术将高维特征组织在一起，利用语义网建立资源的语义关联。原型系统提供了高效率和高准确率的资源检索平台，为医生提供了学习以及交流的平台。实验结果表明，系统的性能得到较大提升。
关键词：医学影像诊断资源平台；语义网；文本处理；影像处理；高维索引

Research on Key Technologies of Medical Image Diagnosis Resource Platform

ZHU Xin-hua[1], ZHAO Da-zhe[1], YU Ya-xin[2], LIU Ji-ren[1]
(1. Software Center, Northeastern University, Shenyang 110179;
2. School of Information Science and Engineering, Northeastern University, Shenyang 110003)

【Abstract】Some key technologies of platform are studied. Text processing technology is used to extract text features. Image processing technology is used to extract grey, texture and shape features. Text and image processing technology are combined to identify and quantify lesions characteristic in medical image. High-dimensional index is used to organize these features. Semantic Web is used to fuse all kinds of resource together. A prototype system provides users with a high accuracy and high efficiency of the resources retrieval platform to study and communicate with each other. Result shows that the performance of system is highly improved after using above technologies.
【Key words】 Medical Image Diagnosis Resource Platform(MIDP); semantic Web; text processing; image processing; high-dimensional index

随着电子学、信息科学的快速发展，医学影像学已成为20 世纪医学领域中知识更新最快的学科之一。医学影像检查在临床医学诊断中发挥着越来越重要的作用，据统计和专家分析，目前医院所使用设备的 1/3 是数字化医学影像设备，医生所做的 80%诊断是依据病人的影像提出的。通过这些医用设备和医院信息化系统的普及使用，在临床的诊断过程中也产生了海量的原始病例信息以及诊断信息。这些未经加工的医疗信息中蕴涵着极为丰富的医学诊断依据和经验描述，从中挖掘和整理出医学诊疗知识，将其有效地管理、组织和利用，是实现医疗诊断现代化的重要手段。本文研究并开发了医学影像诊断资源平台(Medical Image Diagnosis Resource Platform, MIDP)，存储了大量的医学影像诊断相关资源，并提供了灵活的资源检索功能，帮助医务人员，尤其是偏远地区的医生提升基于影像的诊断能力和经验。医学影像诊断相关资源中，除结构化信息外，还包括大量的非结构化多媒体信息如大文本信息，以及医学检查中产生的医学影像等，这些资源在语义上存在着相关性。文献[1-2]分别介绍了文本及影像资源处理的一些基本技术，文献[3-4]介绍了本体及语义网的一些知识。对资源有效的描述和表示是资源能够得到充分利用的前提，MIDP 利用各种媒体间在语义上的相关性和文本及影像处理等技术，构建了针对医学诊断资源的语义网，将这些资源进行整合。

1 MIDP 平台

1.1 平台资源库组成

针对 MIDP 中的医学诊断知识，利用文本处理技术建立关键词索引；对于诊断案例，将与一个历史诊断案例相关的所有信息组织在一起，完整地描述医生一次诊断的过程。

MIDP 的资源组织结构见图 1。

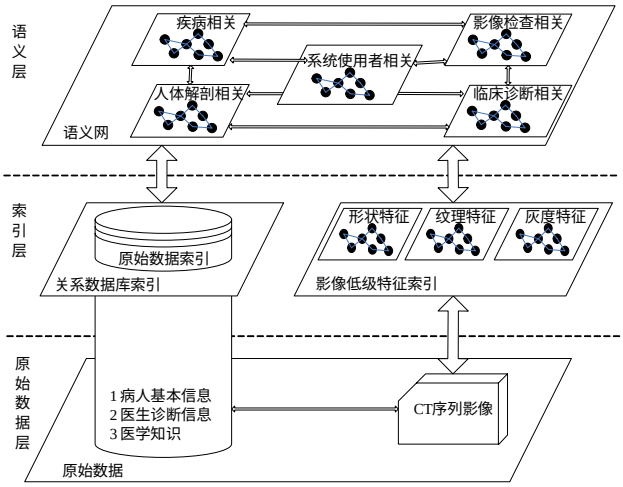


图 1 MIDP资源组织架构

第 1 层为原始数据层，各种结构化数据都存储在关系数

基金项目：国家自然科学基金资助项目“肺癌计算机辅助诊断关键算法研究”(60671050)；辽宁省工程技术研究中心专项计划基金资助项目“远程医疗系统的研究与开发”(辽科发(2005)15 号)
作者简介：朱歆华(1970－)，男，博士研究生，主研方向：并行计算技术，影像处理，人工智能；赵大哲，教授、博士；于亚新，副教授、博士；刘积仁，教授、博士、博士生导师
收稿日期：2007-01-30 **E-mail：**zhuxh@neusoft.com

据库中，医学检查形成的 CT 影像数据以文件形式存储在系统中。第 2 层为索引层，包括从文本和医学影像中抽取的高维特征向量所构建的特征索引，以及数据库索引。第 3 层为语义层，本文基于 ACR 编码(America College of Radiation index)，构建了一个医学影像诊断领域的语义网。语义网层和索引层通过指针建立了关联。原来系统中的各种分散孤立的医学数据，在 MIDP 中形成了完整统一的资源，用户可以根据自己的需要，从各个层面、各个角度查找自己需要的信息和知识。

1.2 资源的特征提取

对于从 PACS 中抽取出结构化数据，直接插入到关系数据库中，对于大文本和医学影像文件等，需要进行文本特征和影像特征的提取。

文本特征是最能代表文本类别和内容的词或短语，为了实现基于词典分词，MIDP 存储了大量中文词汇及其词性，并保存了每个词汇出现的概率。MIDP 还构建了针对医学影像诊断领域的语义网，内容包括与医学影像诊断相关的概念、概念的同义词及近义词、及概念与概念间的关系（如 part-of, instance-of, compose-of 等），语义网可用于理解文本的内容含义，即文本的语义。从文本中提取的语义特征，也将作为语义网实例，补充到语义网中。

算法 1 文本特征提取算法流程

```
输入：Text[]                //原始文本数据
      P_Dictionary     //中文词典指针
      P_Sementic_Web_XML//语义网指针
输出：Word_Vector[]          //分出的词汇
      Concept_Vector[]       //概念词汇
处理流程：
Begin
FOR Text[]中的每段文字
  利用 P_Dictionary 分词，存入 Word_Vector[]
FOR Word_Vector[]中的每个词
  IF Word_Vector 有歧义 OR 该词在 P_Dictionary 未出现
     对 Word_Vector 进行消岐或进行未登录词处理
  End IF
   寻找 Concept_Vector
   建立 Concept_Vector 与 P_Sementic_Web_XML 的链接
   End FOR
End FOR
End
```

本文提取了影像全局特征及影像中病变部位特征。影像全局特征主要包括灰度、纹理、形状等特征。MIDP 还提取了影像中的医学病变部位等医生感兴趣的局部特征，并在影像中自动标注出病变区域。因为病变征象的复杂性，直接利用影像处理算法确定病变位置及病变部位的形态非常困难，但由于在医生诊断描述和诊断结论中，已经包含了对该患者影像中病变部位的描述，因此本文采用了文本与影像交叉参照的方式，利用语义网理解原始数据中的医生诊断描述和诊断结论的语义，然后再利用影像处理技术，到影像中对应的位置确定及标注出病灶，并提取病灶部位的特征向量值。影像病变部位特征提取采用了系统自动提取结合医学专家手工验证的方式。

算法 2 描述了从 CT 序列影像（用 CT 机作一次检查可产生几十或上百幅影像，叫做一个序列）中提取特征的算法流程。

算法 2 影像特征提取算法流程

```
输入：CT_Image_Serials[]          // CT 影像序列
      P_Sementic_Web_XML          //语义网指针
输出：Grey_Feature_Vector[]       //灰度特征
      Texture_Feature_Vector[]    //纹理特征
      Shape_Feature_Vector[]      //形状特征
      Lesions_Feature_Vector[]    //病变部位特征
      Annotated_Image             //标注后的影像
处理流程：
Begin
根据 CT_Image_Serials[]所属案例，找到对应的医生诊断结论
根据 P_Sementic_Web_XML 判断是否属于有病变的影像
确定病变类型及病灶所在位置
FOR CT_Image_Serials[]中的每幅影像
   Extract_Feature(Grey_Feature_Vector[])
   Extract_Feature(Texture_Feature_Vector[])
   Extract_Feature(Shape_Feature_Vector[])
End FOR
IF 该案例无病变 goto End
FOR 病灶所在层数范围中的每幅影像
   增大疑似病变区域
End FOR
确定 CT_Image_Serials[i]为疑似病灶最明显的一层
分割出 CT_Image_Serials[i]中疑似病灶的边缘
Extract_Feature(Lesions_Feature_Vector[])
专家可手工调整病灶分割及特征提取的结果
标注病灶，存入 Annotated_Image
End
```

1.3 资源的检索

从医学影像资源提取出的灰度、纹理、形状等特征都表现为高维特征向量的形式，需要对其进行有效组织和存储才能提高信息检索的性能，MIDP采用了M+树[5]算法建立了高维特征索引，目前主要提供知识查询以及信息查询等两类资源检索模式。

知识查询是指用户输入一些特定的概念， MIDP 首先从资源组织结构中的语义网层面寻找与该概念相关的其他概念，同时向下面的索引层面搜索，最终在原始数据层得到与这些概念相关联的所有医学知识，以及相关的典型诊断案例，并将结果全部返回给用户。

对于信息查询，需用户输入更为明确的查询条件，MIDP 从索引层面开始搜索。因为 MIDP 中的资源主要可分为如病人性别及年龄等结构化信息、医生诊断报告信息、医学影像信息等 3 类信息，因此 MIDP 分别基于这 3 类信息提供信息查询功能。

2 性能分析

MIDP 装载了从某医院获得的 2 000 条医学诊断案例、70 000 余幅相关 CT 影像以及 100 条从各种医学文献中获得的医学诊断知识。系统服务器采用 10 台 PC 服务器集群的方式，每台服务器的硬件配置为：P4 CPU，主频 3.0 GHz，内存 1 GB，硬盘 200 GB，其中一台作为调度服务器。软件环境为：Windows2000 Server, OpenLDAP2.1，数据库采用 Oracle9I。

在诊断案例入库时，分别提取了文本特征、影像全局特征及影像病变部位特征。提取病灶部位特征时，采用了文本和影像处理相结合的方式。

针对基于文本特征、灰度特征、纹理特征、形状特征及影像病变部位特征等 5 种查询方式，分别选用了 10 组查询条

件。基于文本特征的查询条件如表 1 所示。

表 1　基于文本特征的查询条件

查询类型	标记	查询描述
简单查询	Q1	诊断结论中包含“细胞癌” 字样
	Q2	诊断结论中包含“胸膜炎” 字样
	Q3	诊断结论中包含“肺结核” 字样
较复杂查询	Q4	诊断结论中包含“右肺上叶小结节灶”字样
	Q5	诊断结论中包含“右肺下叶结节影”字样
	Q6	诊断结论中包含“左侧胸腔积液” 字样
复杂查询	Q7	诊断描述中包含“右肺上叶尖段见类圆形结节影”字样
	Q8	诊断描述中包含“双侧锁骨下可见肿大淋巴结”字样
	Q9	诊断描述中包含“双肺透过度良好，肺纹理清晰”字样
	Q10	诊断描述中包含“右肺门影增大，右肺下叶起始部见空洞”字样

基于灰度、纹理和形状特征的 10 组查询条件分别采用了头部(2 组)、腹部(2 组)、胸部(2 组)、脊柱(1 组)、颈部(2 组)、上肢(1 组)等部位的例子影像，如图 2。

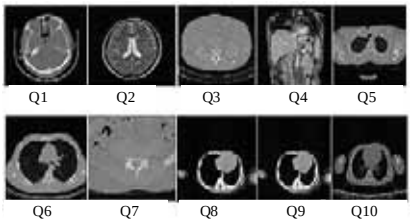


图 2　基于影像全局特征查询用的例子图像

基于影像中病变部位特征，10 组查询条件采用的例子影像，是对病灶部位标注后的影像，而且病灶全部集中在肺部，例子影像形如图 3。

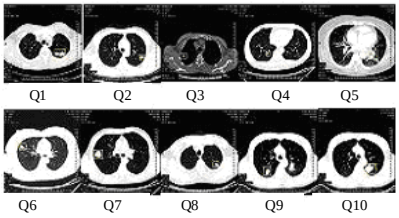


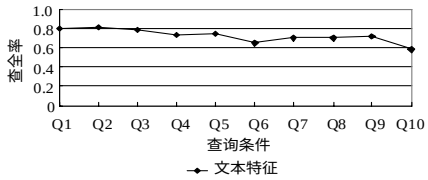
图 3　基于影像病变部位特征查询用的例子图像

利用系统进行查询前，首先对资源库中的文本及影像利用肉眼进行手工分类，分类结果作为评价系统查询性能优劣的依据。本文采用了查全率、查准率作为 MIDP 性能的评价指标，其中查全率、查准率的定义如式(1)和式(2)。

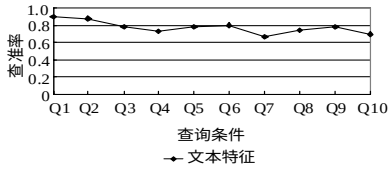
$$\text{查全率}=\frac{\text{有关联的正确检索结果}}{\text{所有有关联的结果}} \tag{1}$$

$$\text{查准率}=\frac{\text{有关联的正确检索结果}}{\text{所有检索到的结果}} \tag{2}$$

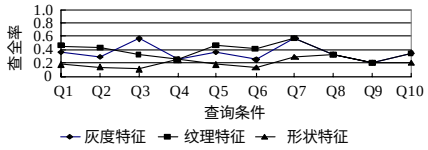
由图 4 可见，基于文本特征的查询图 4(a)、图 4(b)和基于影像病变部位的查询图 4(e)、图 4(f)，查全率及查准率最高，因为这两种查询方式都利用了语义网提供的语义理解技术，所以可以得出结论，在系统中引入语义网技术后，可以提升信息检索的性能。同时，对基于影像全局特征的查询结果图 4(c)、图 4(d)的比较中可以看出，利用累加直方图算法提取出的灰度特征和利用灰度共生矩阵算法提取出的纹理特征的查全率和查准率比较相似，利用 7 个不变距算法提取出的形状特征的查询效果不好，这是由于影像形状特征的提取及表示技术有一定难度造成的。同时，由于头部，胸部及脊柱影像的纹理相对较清晰，因此基于纹理特征的灰度共生矩阵算法更适合于上述 3 个部位的检索。



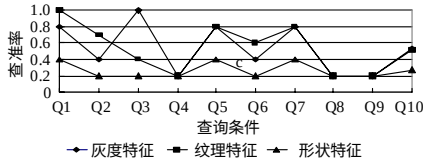
(a)文本特征的查询(1)



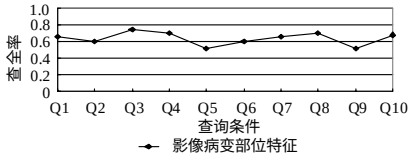
(b)文本特征的查询(2)



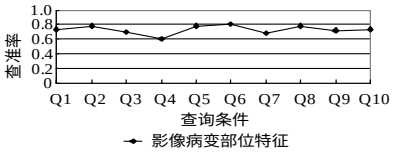
(c)影像全局特征的查询结果



(d)影像全局特征的查询结果



(e)影像病变部位的查询(1)



(f)影像病变部位的查询(2)

图 4　MIDP 的查全率和查准率分析

3　结束语

本文阐述了 MIDP 关键技术及构造过程，MIDP 综合运用了影像处理、文本处理、高维索引等多种技术，并采用了语义网表示出资源间的语义相关性。从性能分析中可以看出，应用这些技术后，可以为用户提供更完善的资源共享服务。

参考文献

[1] 陈　宣, 孔　骏. 基于概率上下文无关文法的句法分析歧义消解新模式[J]. 计算机工程, 2002, 28(2): 126-128.

[2] 田　捷, 包尚联, 周明全. 医学影像处理与分析[M]. 北京: 电子工业出版社, 2003: 234-247.

[3] Berners-Lee T. The Semantic Web and Challenges[Z]. (2003-11-10). http//www.w3.org/2003fTalks/01-sweb-tbll.

[4] OWL Web Ontology Language Reference[Z]. (2004-03-10). http:/lwww. w3.org/TR/2004/REC-owl-ref-20040210/.

[5] Zhou Xiangmin, Wang Guoren, Jeffrey X Y. M+-tree: A New Dynamical Multidimensional Index for Metric Spaces[J]. Australian Computer Science Communications, 2003, 25(2): 161.