

疾病谱数据挖掘与可视化展示系统的设计与实现

李金斌¹ 郝 然¹ 付 磊¹ 王 戎² 赵笑颜³ 尹 岭¹

(1. 中国人民解放军总医院, 北京 100853; 2. 国家卫计委科学技术研究所, 北京 100081;
3. 兰州大学, 甘肃兰州 730000)

摘要: 本文首先对现有人口健康领域科技基础性工作疾病谱相关项目数据资料进行全面梳理、整编, 设计构建了疾病谱数据挖掘与可视化展示系统。然后利用该系统对安徽省阜南县疾病谱数据进行实例分析, 发现县内就诊率不足等问题, 并针对这些问题提出了相关的建议, 经采取相应的措施取得了一定的成效。最后对该系统的建设提出了进一步的改进意见。

关键词: 疾病谱; 数据库; 可视化; 数据挖掘; 数据库建设

中图分类号: G203

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2018.01.015

Design and Implementation of Data Mining and Visualization System for Disease Spectrum

LI Jinbin¹, HAO Ran¹, FU Lei¹, WANG Rong², ZHAO Xiaoyan³, YIN Ling¹

(1.Chinese PLA General Hospital, Beijing, 100853, 2.National Institute of Health Science and technology, Beijing, 100081; 3.Lanzhou University, Lanzhou 730000)

Abstract: Through overall combing and integration of data materials about the fundamentality projects of work disease spectrum in people healthy field, so building special data base being Chinese disease spectrum, and simultaneously reusing data mining and visualization showing, to obtain disease occur law and health care service ability in the region, thus grasping feature of crowd disease spectrum and requirement of medical service in all parts of region, further to promote ability of medical service. This paper researches and designs data mining and visualization system to existing correlation projects of disease spectrum to science and technology fundamentality works. Using the system, it analyzes in application to disease spectrum data in Funan county, Anhui province, so finds the deficiency of visit rate in county, proposes relevant proposal, and adopted relevant measure acquired certain effect. At length, proposes further improved opinions to the system.

Keywords: disease spectrum, database, visualization, data mining, database construction

作者简介: 李金斌 (1989—), 男, 解放军总医院工程师, 主要研究方向: 医学信息学, 数据库建设, 数据挖掘等; 郝然 (1988—), 女, 解放军总医院硕士研究生, 主要研究方向: 医学信息学; 付磊 (1982—), 男, 解放军总医院博士后, 主要研究方向: 医学信息学; 王戎 (1978—), 女, 国家卫生计生委科学技术研究所, 主要研究方向: 医学信息学; 赵笑颜 (1989—), 女, 兰州大学大气科学学院博士研究生, 主要研究方向: 气象医学, 流行病学; 尹岭 (1955—), 男, 博士, 解放军总医院主任医师、教授, 主要研究方向: 医学信息学, 神经内科学 (通讯作者)。

基金项目: 科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”(2013FY110900); 国家人口健康科学数据共享平台专项课题“人口健康平台临床数据资源建设发展策略研究”(2016NCMIZX07)。

收稿日期: 2017年7月14日。

疾病谱是指在整个疾病构成中按疾病患病率（或死亡率等）的高低而排列的顺序，是综合反映医疗单位卫生服务水平的重要指标之一，对了解某地区或医疗机构疾病种类及其变化趋势具有重要意义。因此，疾病谱的数据挖掘与可视化对掌握各地区人群疾病谱特点和医疗服务的需求、研究制定符合各地区实际的医疗卫生政策和疾病防控策略、优化卫生资源配置、促进各地区医疗卫生事业的发展起到重要的决策依据作用。

数据挖掘和可视化展示是近年来医学信息领域研究关注的焦点。数据挖掘技术在统计信息分析系统中的应用，是对数据进行深加工、精加工，成为更直观、更形象的统计信息的重要工具。将统计信息数据和统计分析预测结果以数据表格、统计图表、统计报告等多种可视化方式，通过Web的形式提供给政府或具有访问权限的用户进行实施查询，实现从原始统计数据中挖掘出具有实用价值的信息，实时为政府、相关部门和指定用户提供宏观决策支持^[1]。目前，国外已有不少这方面的成功案例。如应用粗糙集对既往病种病例的医学数据进行数据统计和数据挖掘归纳出诊断规则，可以将人工预测早产的准确率从17%~38%提高到68%~90%。对于患者基本信息，包括年龄、性别、居住地、职业、生活情况等，通过关联规则可以得出具有价值意义的医学关系及诊疗模式^[2-3]。但目前国内对于疾病谱数据库建设和挖掘还较少见。本文将依托现有科技基础性工作疾病谱相关项目数据，开展数据挖掘和可视化系统研究和设计。针对各地区人群诊疗数据选择合适的数据项规范、数据清洗和数据导入，通过对数据的挖掘，建立Web环境下的数据挖掘与可视化展示系统。

1 系统架构与功能需求

本研究开发的疾病谱数据挖掘与可视化展示系统的架构主要采用B/S（浏览器/服务器）三层体系结构，将系统分为表现层、业务层和数据层三层架构，如图1所示。三层架构实现的系统具有更好的灵活性和可重用性。

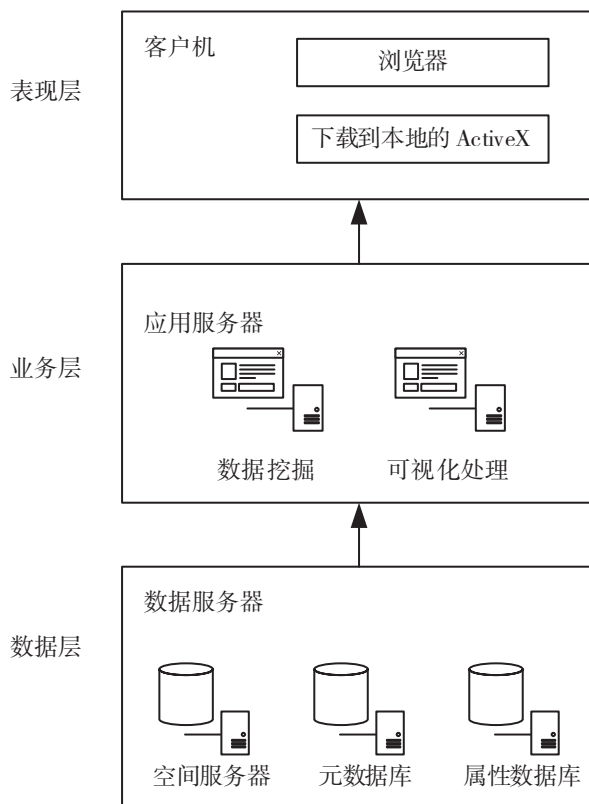


图1 系统逻辑架构图

然而，这里所说的三层是逻辑上的三层，物理上可能为多层。其中，表现层负责地理数据的实时展示，并将一些服务器端实现起来比较困难的功能放在表现层，如用户的交互操作。业务层封装业务的处理逻辑是根据请求生成对数据层的访问，并对得到的数据进行处理生成结果返回给表现层。数据层则由数据库管理系统组成，包括关系数据库和非关系数据库。关系数据库存储属性信息，非关系数据库一般存储地理数据。数据库层独立于应用实现对数据的存储、管理和维护，向业务层提供数据访问的接口^[4]。疾病谱数据挖掘与可视化展示系统具有数据采集系统功能、数据挖掘系统功能和可视化展示系统功能。

1.1 数据采集系统功能

数据可能来源于多个系统，对数据进行脱敏处理。由于某个人的数据可能有多条，应根据唯一识别编码，将其与住院信息、疾病数据及医保数据进行关联，对疾病名称依据ICD-10进行编码，并录入数据库，再根据数据项进行数据

归类。

(1) 数据上传：本系统支持以.xls表格或.sql格式上传数据。

(2) 格式判断：用户按照统一格式上传数据后，系统将对上传的数据进行格式判断，只有符合要求的数据才能成功上传。

(3) 生成数据库表：根据上传的数据自动生成数据库表。

1.2 数据挖掘系统功能

对疾病谱数据资料统计分析的内容包括：人群基本情况分析，疾病和疾病系统发病率排位，系统内疾病发病率排位，并分性别、年龄段对疾病谱构成进行分析比较。对连续多年的数据，还可以分析不同年份疾病谱的变化情况以及不同年份性别、年龄段疾病谱的变化等。在对每个地区进行分析的基础上，对所有的数据进行汇总，分析比较不同地区的疾病谱构成和排位，系统内疾病的构成和排位，并进一步分性别、年龄段对疾病谱构成进行分析比较。

1.3 可视化展示系统功能

可视化系统包括数据上传界面、数据查看界面、图形查看界面、结果导出等子模块。

(1) 数据上传功能：在软件入口上传原始数据文件，导入数据库。

(2) 数据查看功能：可选择地域、年份和民族，生成相应的可视化页面，而在可视化页面中可以多维度地查看数据。

(3) 图形查看功能：可以生成饼图、柱状图、展开图等图形，图中可以拖拽重计算、点击展开、取值对比。

(4) 结果导出功能：可以将挖掘生成的可视化结果导出，在需要的软件中进行展示。

2 数据收集处理与数据库建设

数据库设计方案应该能够保障系统在运行过程中具有较少的数据冗余，并能够有效地降低应用程序与数据库通讯过程中的IO异常发生概率^[5]。为了保障数据库设计的规范性，数据库设计方案应当满足一系列的设计规范。常用的数据库设计

规范为一系列范式，高级别范式对于低级别范式具有包含关系。通常数据库设计应当满足到第三范式。

本文研究的数据主要来源于1999—2012年启动的科技基础性工作专项中人群疾病谱相关项目，包含住院数据、门诊数据、死因监测数据和调查问卷数据等140余万条。根据数据来源和数据项，建立了门急诊病人疾病谱调查数据集、住院病人疾病谱调查数据集、死亡病人疾病谱调查数据集等3个数据集。

在疾病谱调查获取的数据中，有的数据项不完整、不规范，有的是纸质材料，有的项目过多，特别是不同来源的数据之间缺乏联系。为了使数据符合数据挖掘的格式要求，需要对数据进行清洗，剔除数据项不完整、不规范的数据，并以身份证号作为唯一标识，在数据项间建立联系，对数据进行脱敏处理。通过数据处理系统，将不同来源、不同格式的数据进行数字化、规范化，形成可用于数据挖掘的人群疾病谱数据库。

MySQL是关系型数据库管理系统应用软件之一，由瑞典MySQLAB公司开发，目前属于Oracle公司。本文将采用关系模型的MySQL数据库系统，对已经导入数据库中的经过预处理的数据通过数据库表的形式进行保存，这就相当于数据仓库式的存储。SQL是关系数据库的标准访问语言，具有体积小、速度快、总体拥有成本低、开源等优点，而MySQL对其提供了很高的支持度。因此，采用MySQL数据库系统，不仅提高了运行速度，而且提高了灵活性。

3 数据挖掘与可视化展示

3.1 数据挖掘

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的人们事先不知道但又是潜在有用的信息和知识的过程。而关联规则表示数据库中一组对象之间某种关联关系的规则。关联规则挖掘就是通过关联分析找出数据库中隐藏的关联，利用这些关联规则可以根据已知情况对未知问题进行推

测^[6]。将数据挖掘技术应用到医学信息数据库中，可以发现其中的疾病谱规则和模式^[7]。

针对疾病谱数据的数据量大、数据复杂度较低等特点，本文采用K-means算法，将相似性大的数据点聚于同一个簇，把具有较大差异性的数据放在不同的簇，相似性和差异性由具体规定条件加以约束。通过数据统计和数据挖掘，归纳出疾病发生和诊疗情况的基本规则，与患者基本信息，包括性别、年龄、居住地、家庭情况等，通过关联规则得出具有一定价值意义的医学关系及规律模式^[8]。以此展示了医疗特点和疾病规律适用的数据挖掘模型，并利用数据挖掘技术实现了相应的统计分析功能^[9]（图2）。

3.2 可视化展示

数据挖掘流程通常集成了多个挖掘算法，传统上挖掘算法的运行结果通常以文本的形式显示，内容复杂且不利于理解。而数据挖掘结果可视化则是以图形化的方式直观展示各类算法的运行结果，方便用户更加快速直接地理解结果，发现问题，指导后续挖掘工作的调整。

数据可视化展示系统是基于二维矩阵的可视化技术建立的。该方法利用一个二维矩阵的行和列分别表示规则的前项和后项，并在对应的矩阵单元画图，可以是柱状图或条形图等。不同的图形元素（如颜色或高度）可以用来描述关联规则的不同参数，如规则的支持度和置信度。不同颜色柱状图的高度分别表示支持度和置信度的置信度取值。二维矩阵法的优点是易于可视化一对一布尔关系的关联规则；局限性在于当大量的关联规则需要可视化时，后面的图形会被前面的图形遮蔽而不容易观察。这样的数据可视化展示系统可以根据用户的需求进行多类型字段进行复合检索、查询，对检索结果进行数据的分类统计，并根据用户需求采用相应的算法，生成Json格式的数据，提交给前端，生成可视化图表（图3，图4）。

4 实例分析

疾病谱数据库收集了安徽省阜南县14个乡镇2013—2016年门诊、住院数据。利用疾病谱

The screenshot displays the '疾病谱数据管理' (Disease Spectrum Data Management) system interface. The top navigation bar includes '业务管理', '系统设置', '日志查询', and '代码生成'. The main area shows a '数据查询' (Data Query) section with a search bar and filters. The search criteria are set to '住院表' (Inpatient Table) with a filter for '性别' (Gender) set to '男' (Male). The results are displayed in a table with columns: '序号' (Serial Number), '合作医疗证号' (Cooperative Medical Certificate Number), '性别' (Gender), '年龄' (Age), and '住院费' (Inpatient Fee). The table lists 12 records.

序号	合作医疗证号	性别	年龄	住院费
1	0508040017	男	46	0
2	0509020059	男	61	0
3	0305010003	男	72	69
4	0305080095	男	73	133
5	0407060007	男	56	115
6	0509030020	男	34	304
7	0601020041	男	44	240
8	0733020301	男	18	171
9	1004020013	男	54	92
10	1122020014	男	35	95
11	0509020059	男	61	120
12	0201050065	男	52	350

Copyright © 2012-2015 疾病谱数据管理 - Powered By V1.0

图2 疾病谱数据库数据检索功能



图 3 疾病谱数据地理信息展示

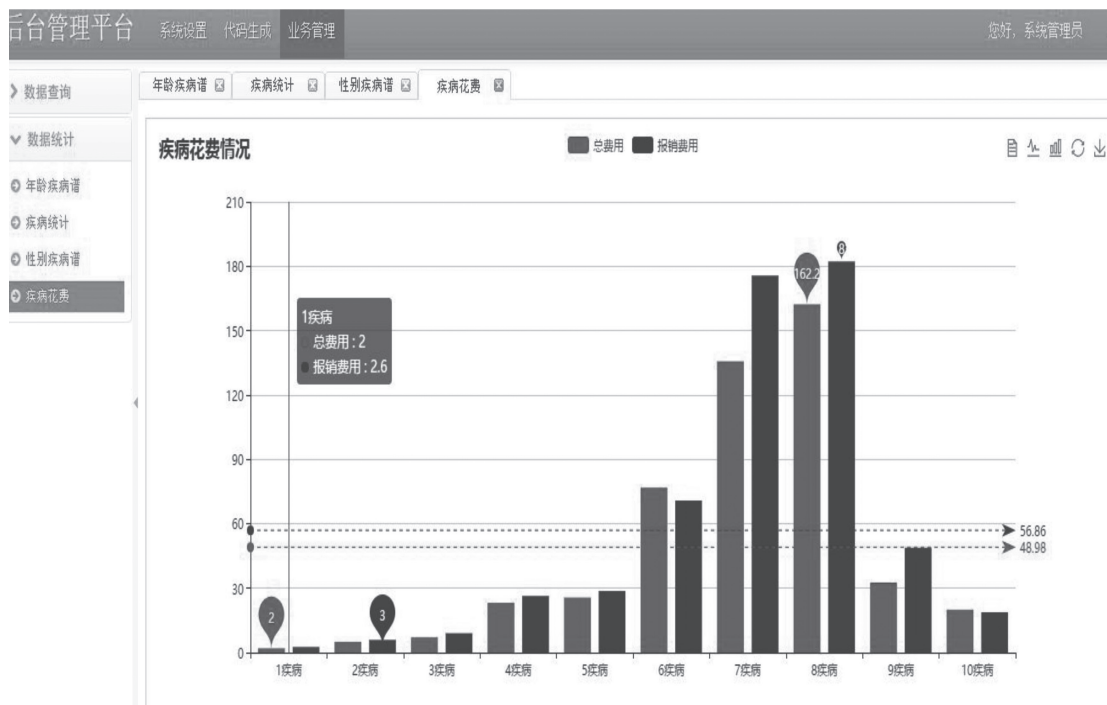


图 4 疾病负担情况查询可视化展示

数据库对疾病谱数据进行处理分析，得到阜南县居民疾病谱和疾病负担以及医疗服务现状和需求，为地区医疗卫生事业的发展提供了决策依据。

疾病谱数据显示，2014—2015 年安徽省阜

南县县内的就诊率不足 60%，大量的患者不得不走出县域就诊。而对于同样的手术（如阑尾炎手术），同样的设备，同样的技术，除报销外的自负部分，省级医院次均需要近 6000 元，市级医院次均需要 4500 元，县医院次均需要 1500 元，

而乡镇卫生院次均只需要 100 元。可见走出县城就诊极大地增加了老百姓的医疗费用。如何提高县域的医疗服务能力, 解决百姓的看病问题成为健康阜南县的首要任务。

综合居民疾病谱和阜南县医疗服务能力现状, 提出了阜南县医共体模式, 建立了阜南县医共体。在医共体内, 按照分工, 县级医院看大病、乡镇卫生院治小病、村卫生室防未病。具体而言, 县级医院承担着 100 种以上的大病诊治工作, 从而降低了阜南县患者医疗费用支出。

在阜南县, 县人民医院重点加强肿瘤、心血管、重症等科室建设, 组建肿瘤、心脑血管等六大中心; 县中医院重点加强中医、骨科等科室建设, 组建中医治疗中心; 县三院重点加强康复、老年病等科室建设, 组建康复治疗中心; 妇幼保健院重点加强妇产科、儿科的建设, 组建妇儿病治疗中心。这样就提高了大病诊治能力。

在安徽省阜南县通过应用示范, 提升了阜南县基层卫生决策和医疗服务整体水平, 实现了全面提升医疗服务水平, 满足了阜南居民医疗服务和慢病管理需求。据疾病谱数据统计, 阜南县县外转诊率 2016 年较 2015 年改革前下降了 12.98%, 而县外住院为 4.1 万人次, 占比为 22.68%, 较改革前下降 4.7 个百分点从而降低了群众医疗负担。

5 结语与讨论

针对人口健康领域基础性工作疾病谱相关项目数据资料的全面整编, 形成疾病谱调查专题数据库, 并设计实现了疾病谱数据挖掘与可视化展示系统。该系统性能可靠, 可通过多类型字段进行复合检索、查询, 可进行数据的分类统计, 实现通过性别、年龄、费用、病种等的统计结果的可视化展示, 获得地区居民疾病谱、疾病负担的具体数据, 得到地区疾病发生规律和医疗卫生服务能力现状, 结合当地实际情况, 为政府医疗机构提出医疗卫生事业的发展决策建议和医疗卫生机构学科建设和人才培养计划, 大幅度提升全市

医疗卫生服务能力, 满足人民群众基本医疗卫生服务需求和多样化、多层次健康需求。但是, 在设计过程中, 有一些需求还没有考虑到, 实现的功能也还有很多不完善的地方, 随着数据量的不断增大, 分析需求的不断增加, 还应考虑系统的扩展性, 对系统功能进行扩充。今后, 应在以下两个方面进行完善: 一是目前项目研究数据集中在单个地区其研究多数为独立开展, 数据来源不具有广泛性, 只能得出单独地区疾病谱和结论, 尚未得到区域性的人群疾病谱数据的支持, 需要更多的疾病谱数据完善数据库系统。二是目前的系统对数据挖掘模型的设计还仅考虑了一些基本需要, 尚需对数据进行更加深入的数据挖掘, 如从卫生经济学角度对疾病谱数据进行挖掘, 可以得出某地区居民疾病发病率和疾病负担的关系, 为当地提出人才培养和疾病防控等方面的建议, 能够为当地优化医疗卫生资源合理配置。

参考文献

- [1] 肇恒宇. 数据挖掘技术在医学领域中的应用[J]. 中国科技信息, 2009(15): 129-130.
- [2] 姚远, 刘丽华. 军队慢性疾病诊疗数据库与数据挖掘研究综述[J]. 中国医院, 2014, 18(9): 74-76.
- [3] 黄科军, 耿生玲. 统计信息的数据挖掘与可视化分析系统[J]. 硅谷, 2010(9): 40-41.
- [4] 刘婉姮. 少数民族人群诊疗数据分析与可视化展示系统的设计与实现[D]. 天津: 南开大学. 2015.
- [5] 段远志. 数据库设计方法[J]. 南昌教育学院学报, 2005(4): 84-86, 89.
- [6] IOANNIS K, OLGA T, ATHANASIOS S, et al. Machine learning and data mining methods in diabetes research[J]. Computational and Structural Biotechnology Journal, 2016.
- [7] 于长春, 贺佳, 范思昌, 等. 数据挖掘技术在医学领域中的应用[J]. 第二军医大学学报, 2003(11): 1250-1252. DOI: 10.16781/j.0258-879x.2003.11.034.
- [8] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained k-means clustering with background knowledge[C]//ICML. 2001, 1: 577-584.
- [9] 于洋. 数据挖掘可视化技术的研究与应用[D]. 长春: 吉林大学, 2008.