农民工医疗健康信息分析系统的维度建模设计①

王 超

(天津大学 管理与经济学部, 天津 300072)

摘 要: 农民工医疗健康问题是政府长期关注的难题之一, 农民工医疗健康信息分析系统旨在对农民工医疗健康信息数据进行统计与挖掘, 辅助政府决策. 基于数据仓库中的维度建模理论, 结合农民工医疗保健需求及卫生服务活动特点, 建立面向数据分析的农民工医疗健康信息的多维数据模型, 对数据分析及挖掘的应用及方法进行初步设计, 为农民工医疗健康信息数据分析提供基础, 并为政府相关部门提供参考.

关键词: 维度建模; 数据仓库; 数据挖掘; 农民工; 医疗

Dimensional Modeling of the Migrant Workers Medical Data Analysis System

WANG Chao

(School of Data Analysis System, Tianjin University, Tianjin 300072, China)

Abstract: The Migrant Workers Medical Data Analysis System aims at analyzing medical care data of migrant workers based on data warehouse and OLAP technology, in order to assist decision making for government. Data modeling is one of the most important steps of building data warehouse. Therefore, this paper designed a multi-dimensional model for the system based on dimensional modeling theory, and provided the elementary design of data mining analysis.

Key words: dimensional modeling; data warehouse; data mining; migrant worker; medical care

1 引言

农民工主要是指户籍仍在农村, 进城务工和在当地或异地从事非农产业的劳动者[1], 是我国特有的城乡二元体制的产物. 随着我国工业化、城镇化进程的加快, 农民工的数量逐年上升, 2011 年全国农民工总量已达到 2.5 亿人[2]. 与此同时, 农民工的医疗健康问题也越来越受到关注. 农民工医疗健康所涵盖的数据量相当庞大并且日益增长, 合理组织并利用这些数据, 从中发现有用的规律并预测相关问题的发展趋势, 可以帮助政府和相关机构从宏观角度把握农民工医疗健康问题, 作出相应调整, 从而使农民工享受到更完善的医疗服务.

农民工医疗健康数据具有海量性与异构性, 采用数据仓库技术是合理的选择, 结合 OLAP 技术和数据挖掘技术进行数据分析、挖掘和展现, 可实现多视角的深层透视与分析. 数据建模是数据仓库设计中最重要的步骤之一, 它决定了数据存取查询的质量和效率, 建立一个可靠的、可伸缩的、以及可维护的数据仓库维度数据模型至关重要[3]. 数据仓库的数据建模方法主要有 Inmon[4]倡导的第三范式建模方法和 Kimball[5]提出的维度建模方法等. 其中维度建模方法可以更好地展现多维数据关系, 并具有易于理解、访问效率较高等特点, 因此采用维度建模方法对农民工医疗健康信息进行数据建模, 并对相关数据挖掘方法进行设计.

2 维度建模概述

维度建模(Dimensional Modeling)技术是数据仓库数据建模的主要方法之一, 由 Ralph Kimball 提出[5], 是一种将数据模型概念化和形象化为一组可用一般商业概念度量的技术[3], 它从业务需求出发, 用维度和事实描述业务中的分析对象. 事实表(Fact Table)是维度模型中的基本表, 由度量值和连接到维度表的外关键字组成. 度量值又被称作“事实”, 用于描述业务性能, 是数据分析的核心. 维度表(Dimension Table)包含对业务的

① 基金项目:国家科技支撑计划(2011BAH15B04)
收稿时间:2012-09-25;收到修改稿时间:2012-11-02

描述，由维度属性组成. 维度属性是查询的约束条件，是分析切割的依据，为用户提供了使用数据仓库的接口. 事实表和维度表通过外关键字(Foreign Key, FK)连接，事实表中的外关键字子集构成其自身的复合主关键字(Primary Key, PK). 事实表与维度表的融合构成了维度模型的基本框架. 维度模型是一种多维数据模型，主要有三种模式: 星型模式、雪花模式、事实星座. 星型模式由一个事实表和围绕它的一组维度表组成，每个维度仅拥有一个维度表，存在大量冗余，是一种非规范化的模式，效率较高. 雪花模式将维度表中的属性以规范化的形式进一步分解到附属表中，可以表示出属性间的层级关系，减少了冗余且利于维护，但增加的表连接操作带来了效率的降低[6]. 事实星座是由多个事实表共享维度组成的更为复杂的模型.

基于 Kimball 的理论[7]，采用维度建模方法实现的农民工医疗健康信息分析系统数据仓库结构如图 1.

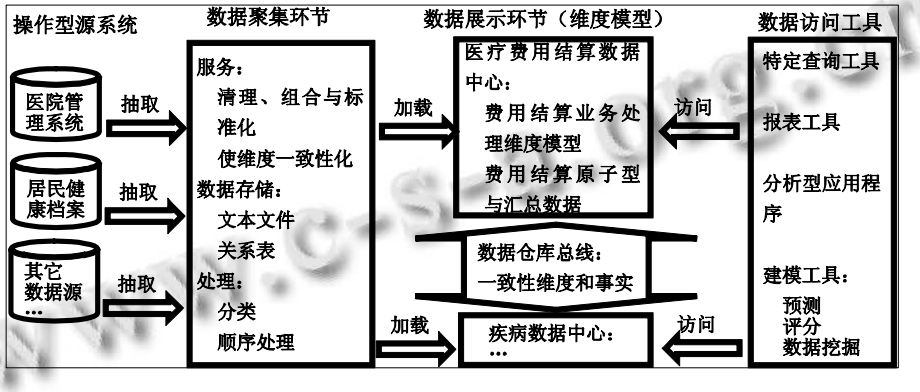

图 1　农民工医疗健康信息分析系统数据仓库的基本结构

3　农民工医疗健康信息分析系统需求分析

农民工医疗健康分析系统旨在对医疗数据进行统计分析及数据挖掘，发现其中有价值的信息，以辅助农民工医疗监管及政策改革. 系统面向数据分析，主要考虑就诊规模分析、疾病风险分析、医疗费用支付、医疗保险类型分析等方面，需求包括以下两个层次.

(1) 数据汇总与统计分析. 为政府、卫生服务管理机构、医疗服务组织、研究机构等组织提供其感兴趣的数据分析结果，为疾病防治和医疗监管提供依据. 典型应用如: 按地区、职业、年龄等人口统计学特征查询某种疾病的发病情况; 按地区、疾病类型等条件查询农民工对医疗费用的承担情况; 查询医疗费用的构成情况，探究其分布的合理性.

(2) 数据深度挖掘分析，发现有用的规律及趋势，为公共卫生、疾病监测等工作提供依据. 典型应用如: 按地区、职业等特征对患者分布和疾病发生风险进行关联规则挖掘分析; 依据日期维度建立某地区某疾病发生的预测模型等.

4　农民工医疗健康信息分析系统维度建模

4.1 建模对象业务选取

业务处理过程是建模的对象，一般是由数据源系统所支持的自然业务过程. 医疗服务过程包括医院就诊、社区卫生服务、疾病预防控制、妇幼保健、卫生监督等. 基于需求分析，将业务主题确定为农民工疾病风险主题、医疗费用主题等几方面，选取具代表性的医院就诊活动及费用结算过程两个业务作为主要业务处理流程，建立疾病事实表和费用结算事实表.

粒度确定了事实表中每一行记录所表达信息的具体含义，决定了维度与事实的关联关系，同时也决定了数据仓库中数据量的大小以及数据查询活动所能回答问题的细节程度. 为了提高查询分析的灵活性，通常需要建立基于原子粒度的原子事实表，同时可配合建立聚集事实表[7]. 选取一位病人所诊断出的一项疾病作为疾病事实表的粒度，选取一次就诊活动中的费用结算清单作为费用结算事实表的粒度.

4.2 维度表设计

维度的属性选取主要依据卫生部发布的卫生信息数据元标准. 其中日期维、个人基本信息维等共享维度设计为一致性维度.

(1) 日期维. 用于描述事实发生的时间或聚集依据. 事实粒度中将时间维度的最小粒度设置为日，概念层次选为年、月、日，为不同级别的数据汇总查询

提供支持. 为方便从日期特征角度对就诊及疾病风险进行分析, 加入季、星期、节假日指示符等属性. 日期维表主要属性集为: 日期关键字(PK), 日期, 日期完整描述, 年, 季, 月, 日, 星期, 节假日指示符).

(2) 个人基本信息维. 用于描述农民工个人的人口统计学特征, 为数据汇总提供依据. 主要属性集为: 个人基本信息关键字(PK), 姓名, 性别, 身份证号码, 出生日期, 年龄, 民族, 婚姻状况, 户籍类型, 户籍所在地, 常住地, 常住地址户籍标志, 职业类别代码, 职业类别名称, 职称代码, 家庭人均年收入, 参加工作日期, 药物过敏史标志, 药物过敏源, 主要医疗保险支付方式, 次要医疗保险支付方式. 常住地、职业、收入、医疗保险类型等属性可能发生变化, 该维度属于渐变维度(slowly changing dimensions, SCD), 为便于变化跟踪, 建立为 SCD2 型维度.

(3) 卫生机构维. 用于描述医疗卫生服务活动的提供方, 地区信息选取省(自治区、直辖市)、市、区(县)三个层次. 主要属性集为: 卫生机构关键字(PK), 卫生机构名称, 卫生机构代码, 卫生机构类型, 所在省(自治区、直辖市), 所在市, 所在区(县), 行政区划代码. 其中卫生机构类型主要包括医院、门诊、妇幼保健院、专科疾病防治院等.

(4) 疾病维. 用于描述在就诊过程的医学诊断结果. 属性集为: 诊断关键字(PK), 疾病诊断代码(ICD-10 代码), 疾病诊断名称, 疾病类型, 附加描述.

(5) 诊断组维. 在医院就诊活动中, 一次就诊活动和费用结算活动可以包含一个主要诊断和多个其它诊断, 因此疾病诊断维度是一个多值维度. 诊断个数为不确定值, 采取桥接表的方式连接疾病维与事实表. 诊断组桥接表属性集为: 诊断组关键字(PK, FK), 诊断关键字(PK, FK), 费用权重因子. 其中费用权重因子为该疾病占诊断组的费用比例, 用以支持疾病费用汇总.

(6) 诊断检查技术维与桥接表. 诊断检查技术维用于描述诊断过程中采取的临床辅助检查技术. 属性集为: 诊断检查技术关键字(PK), 诊断检查技术代码, 诊断检查技术名称, 附加描述. 诊断检查技术是多值维度, 与诊断维类似, 建立诊断检查技术组桥接表: 诊断检查技术组关键字(PK, FK), 诊断检查技术关键字(PK, FK).

(7) 治疗维与桥接表. 治疗维用于描述医疗服务过程中采取的治疗手段. 主要属性集为: 治疗关键字(PK), 手术及操作代码、手术及操作名称、其它治疗类型名称, 治疗描述. 治疗是多值维度, 同时建立治疗组桥接表: 治疗组关键字(PK, FK), 治疗关键字(PK, FK).

(8) 治疗结果维. 治疗结果维用于描述治疗的结果状态. 主要属性集为: 治疗结果关键字(PK), 出院情况, 诊断符合情况. 其中出院情况包括治愈、好转、转院、死亡等层次, 诊断符合情况包括符合、不符合等层次.

(9) 医疗费用支付方式维. 用于描述农民工医疗保险报销方式. 主要属性集为: 医疗费用支付方式关键字(PK), 医疗费用支付方式代码, 医疗费用支付方式名称, 其它描述. 其中农民工可选的医疗费用支付方式主要有: 城镇职工基本医疗保险、新型农村合作医疗保险、贫困救助、商业医疗保险、全公费、全自费、其它社会保险、其它等.

(10) 医疗费用结算方式维. 医疗费用结算方式包括: 现金、支票、汇款存款、内部转账、单位记账、账户金、统筹金、银行卡等. 其主要属性集为: 医疗费用结算方式关键字(PK), 医疗费用结算方式代码, 医疗费用结算方式名称, 其它描述.

4.3 事实表设计

根据业务选取, 确定如下事实表.

(1) 疾病事实表. 疾病事实表以病人就诊活动为依据, 以一项疾病诊断为粒度, 记录病人患病史, 主要用于农民工疾病风险分析. 事实主要包括住院标志和手术标志、住院天数、累计计数等. 其中住院标志和手术标志属于不可加型事实, 累计计数恒为 1, 作为可加型事实方便汇总统计. 维度外关键字包括: 个人基本信息关键字、疾病关键字、治疗结果关键字、诊断日期关键字、住院日期关键字、出院日期关键字、医疗机构关键字, 另外加入病案号作为退化维度, 用以标识和汇总. 其星型结构如图 2 所示.

(2) 费用结算事实表. 费用结算事实表以一次医疗结算单据为依据, 包含与费用发生相关的各项信息. 度量值包括: 总费用、可报销费用、个人承担费用、各项分类费用等. 分类费用主要包括住院总费用、门诊总费用、诊断费用、化验费用、手术费用、非手术治疗费用、康复费用、西药费用、中成药费用、中草药费用、中医费用、血液和血液制品费用、耗材费用、其它费用等. 将各分类费用作为事实属性置于一张费

用事实表中可减少事实表行数, 同时便于汇总查询. 维度外关键字包括就诊日期关键字、结算日期关键字、个人基本信息关键字、卫生机构关键字、诊断组关键字、诊断检查技术关键字、治疗组关键字、治疗结果关键字、医疗费用支付方式关键字、其它医疗费用支付方式关键字、医疗费用结算方式关键字、病案号(退化维度). 其星型结构如图 3.

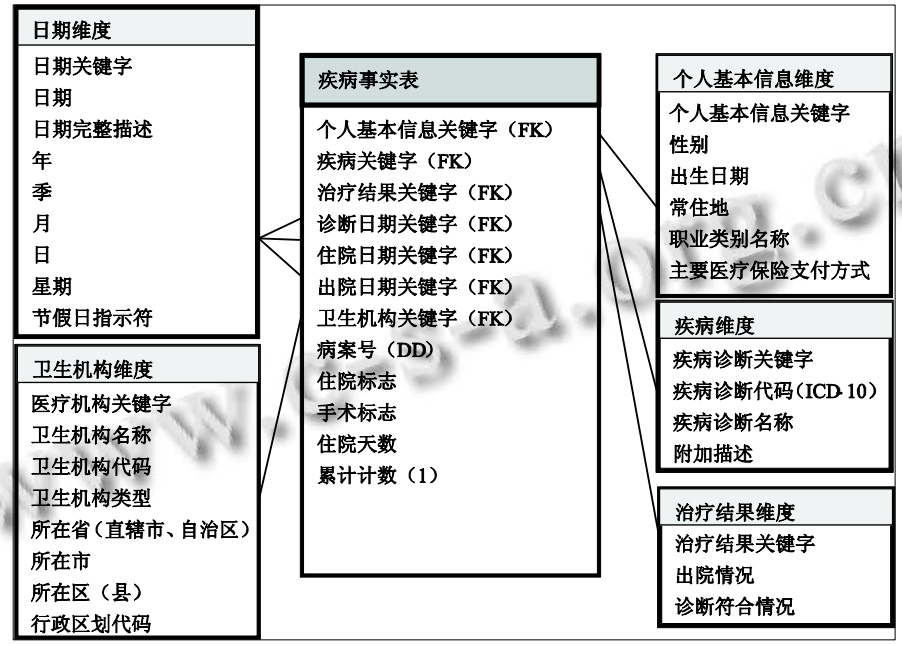

图 2　疾病事实表星型结构

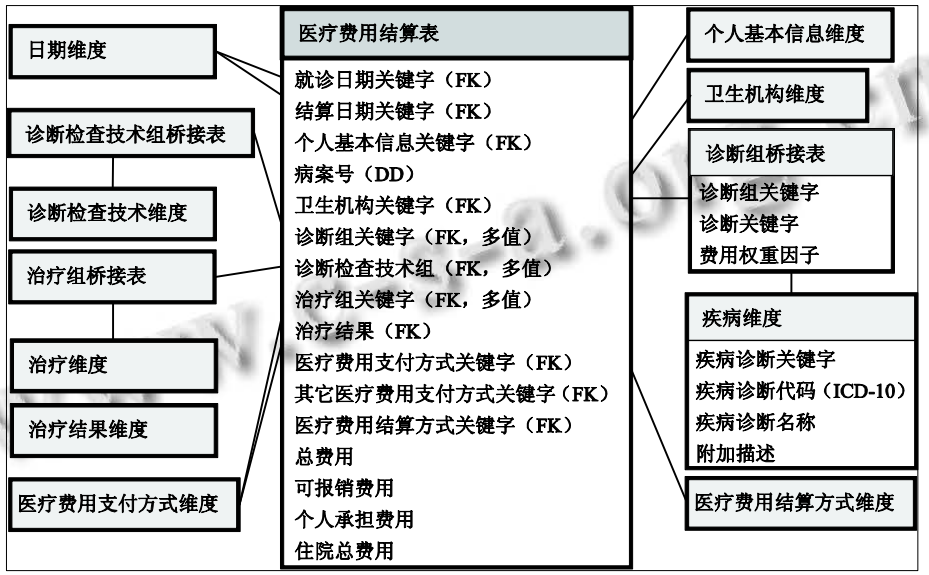

图 3　费用结算事实表星型结构

5　基于维度模型的数据分析及挖掘设计

该维度模型可以满足数据分析与数据挖掘的基本需求, 主要应用如下.

5.1 农民工疾病风险分析

(1) 就诊规模分析. 以费用结算事实表和个人基本信息维为中心, 分析就医农民工的年龄、职业、性

别等特征分布; 沿日期维度上卷, 以月份或年份为依据分析就诊规模的趋势变化; 以疾病为条件分析各疾病的就诊人数构成及趋势变化.

(2) 农民工疾病风险的关联规则分析. 以疾病事实表为中心, 对职业、地区属性和疾病类型进行关联规则挖掘, 分析职业、地区、年龄等人口信息与某疾病发生的关系; 同时结合治疗结果属性统计各疾病的平均住院时间、治愈率、死亡率等指标, 并在疾病维度上沿疾病类型进行上卷和下钻操作, 从而发现影响农民工疾病发生的风险因素.

(3) 农民工疾病风险的趋势分析. 以疾病事实表为中心, 沿诊断日期维度进行上卷, 以月份或年份为汇总依据对疾病发生的趋势进行分析和预测, 为疾病监管与防治工作提供依据.

5.2 农民工医疗费用结算数据分析

(1) 医疗费用分布统计. 以费用结算事实表为中心, 选取总费用、个人承担费用及各项分类费用为统计度量值, 依据地区、职业类型、年龄、性别、疾病类型等维度属性进行切块, 对农民工就诊过程中的医疗费用的分布进行统计, 并依据年度做出费用趋势变化分析, 作为地方医疗保障政策调整的依据.

(2) 异常费用识别. 采用聚类分析方法识别离群点可以分析出异常费用. 基于费用结算事实表, 选取疾病类型、年龄、地区、住院总费用、各项分类费用、报销费用等作为聚类变量, 进行聚类分析, 识别其中离群点, 对不合理的医疗费用支出和报销加以分析和监控, 辅助医疗费用增长控制及报销监督审查.

5.3 医疗保险类型分析

(1) 参保规模统计. 基于个人基本信息维度统计各年度各地农民工参加医疗保险的参保率、各类医疗保险的参保分布情况等, 对参保规模进行趋势分析, 对各年度、各地农民工医疗保险覆盖率进行分析.

(2) 各保险资金支出统计. 基于费用结算事实表统计各类医疗保险在费用结算时的支付情况, 按年度、地域、保险类型汇总, 考察各类保险的资金使用情况.

(3) 参保类型分析. 基于个人基本信息维度对各保险类型进行构成分析, 统计各年度每类保险参保人数的变化率, 分析各保险类型的新增人数、退保人数、由其它险种转入人数、转出为其它险种人数等, 对各类保险进行比较.

6 结语

目前关于面向数据分析的医疗数据建模研究较为缺乏, 农民工医疗健康信息分析系统以数据分析为核心需求, 将农民工个人信息、疾病信息与费用信息相结合, 构建更为全面、适于数据分析的数据仓库维度模型; 并提出统计分析和数据挖掘的方案设计, 为进一步完善农民工医疗健康信息分析系统提供基础.

参考文献

1 国务院研究室课题组.中国农民工调研报告.北京:中国言实出版社,2006:1–2.

2 国家统计局.2011 年我国农民工调查监测报告. [2012-07-21]. http://www.stats.gov.cn/tjfx/fxbg/t20120427_402801903.htm.

3 Kimball R, Ross M. The Data Warehouse Lifecycle Toolkit, 2nd Edition. New York: Wiley Computer Publishing, 2008: 3–9.

4 Inmon WH. Building the Data Warehouse. New York: Wiley Computer Publishing, 1992: 52–59.

5 Kimball R, Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. New York: Wiley Computer Publishing, 1996: 1–27.

6 Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record, 1997,26(1):65−74.

7 Kimball R, Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2nd Edition. New York: Wiley Computer Publishing, 2002: 6–8.