文章编号:1671-4598(2017)12-0191-04 DOI:10.16526/j.cnki.11—4762/tp.2017.12.050 中图分类号:R319 文献标识码:A

基于医院数据中心的临床全视图构建方法研究

孟 亚,严 健,黄优杰,刘鹏远
(上海市徐汇区大华医院,上海 200237)

摘要:为了提高医院临床管理的精细化程度,更好地满足临床诊疗、科研和医疗质量的需求,需要构建基于医院数据中心的临床全视图系统;当前的临床全视图构建方法,是以数据分块存储的方式进行构建,导致临床医护人员无法获得完整有效的医疗数据,存在临床数据分散、数据不完整等问题;为此,提出一种基于医院数据中心的Paxos算法临床全视图构建方法;仿真实验结果证明,所提方法可以将医疗数据有效地应用到医务人员的临床工作中,达到数据资源利用最大化,帮助医院实现医疗信息化改进和服务创新,使各个异构系统之间的数据进行交互,实现了医疗数据共享,为临床工作的发展提供了可用工具;基于医院数据中心构建的临床全视图具有全面、精准、可共享的特点,在未来医院数据中心控制系统的发展中具有重要作用。

关键词:医院数据中心;全视图构建;方法研究

Research on Construction Method of Clinical View Based on Hospital Data Center

Meng Ya, Yan Jian, Huang Youjie, Liu Pengyuan
(Shanghai Dahua Hospital, Shanghai 200237, China)

Abstract: In order to improve the refinement of clinical management and better meet the needs of clinical diagnosis and research and medical quality, it is necessary to build a clinical whole view system based on hospital data center. The construction method of current clinical view, is based on data block storage mode, leading to clinical medical staff difficult to obtain complete medical data effectively, clinical data scattered and incomplete data problems. Therefore, a method of constructing clinical full view of Paxos algorithm based on hospital data center is proposed. The simulation results show that the proposed method can be effectively applied to medical data in the clinical work of medical personnel, to maximize the use of the data resources, help hospital to realize medical information service innovation and improvement, the interaction between heterogeneous system data, the implementation of medical data sharing, provide the tools available for the development of clinical work. The clinical full view based on hospital data center has the characteristics of comprehensive, accurate and sharable. It plays an important role in the development of hospital data center control system in the future.

Keywords: hospital data center; full view build; methods to study

0 引言

目前,随着医院信息系统的全面发展,在临床工作中产生了海量医疗数据,例如门急诊数据、手术数据、住院数据等[1]。医疗数据的用处很多,它不仅可以为医疗技术的提高提供辅助作用,而且还可以帮助医院进行管理创新[2]。但由于医疗数据错综复杂,很难将所有的相关数据集合到一个页面展示,尤其是在临床方面,因此医院临床全视图的构建成为了当今HIT业界的热点问题[3-4]。医疗数据具有规模庞大,数据交互性不强等特点,多数医院临床全视图构建方法无法对医院临床数据进行准确稳定地构建,导致基于医院数据中心的临床全视图系统构建时,经常出现恶意数据混淆,可利用数据丢失,数据查找结果不明确等问题[5]。在这种情况下,如何提高基于医院数据中心临床全视图构建准确度,增加全视图构建质量,成为了该领域亟待解决的问题[6]。而利用Paxos算法进行基于医院数据中心的临床全视图构建的方法,不仅可以对临床全视图进行全面,高效地构建,而且也是解决上述问题的有效途径,受到了医疗信息化方面专家的关注和深度钻研,同时也出现了很多好的方法[7]。

文献[8]提出了一种基于本体论的医院临床全视图构建方法。该方法首先对医院信息数据网络的生成进行深入研究,然后利用全局本体对医院临床全视图的构建进行透彻分析,最后结合实例,完成对医院临床全视图的构建。该方法运行起来很简单,但是存在构建准确率低的问题。文献[9]提出了一种基于XML中间件技术的医院临床全视图构建方法。该方法首先在医院数据集成的基础上,不局限于原有数据,对数据源的范围进行拓宽,然后利用中间件技术独立开发的性能,使数据彼此间固定接口并进行交互,明确各自的功能,最后以数据源作为单位,构建医院临床全视图。该方法用时较短,但是对数据源范围的拓宽过程导致了恶意数据较多的问题。文献[10]提出了一种基于Datalog规则的医院临床全视图构建方法。该方法首先对医院数据集成结构进行分析,然后利用

收稿日期:2017-05-10; **修回日期**:2017-05-22。

作者简介:孟 亚(1981-),男,上海人,硕士研究生,高级工程师,主要从事医院信息化管理与建设相关工作方向的研究。

通讯作者:严 健(1958-),男,江苏人,硕士研究生,主任医师,主要从事医院管理方向的研究。

Datalog 规则将医院临床全视图描述出来，最后将数据集成分为自上而下和自下而上两类，并对医院数据全视图的集成部分进行讨论。该方法对医院临床全视图构建的比较全面，但是存在耗时较长的问题。

针对上述产生的问题，提出一种基于医院数据中心的 Paxos 算法临床全视图构建方法。仿真实验证明，所提方法可以准确地对医院临床全视图进行构建。

1 基于医院数据中心 Paxos 算法临床全视图的建设方法

1.1 医院数据中心建设

1.1.1 医院数据清洗

以 2.1 中的信息为依据，利用扩展树状知识库对医院数据中心的数据进行清洗。其清洗过程为：首先根据医院实际情况的需求，获得清洗的相关知识和该知识对应的原子知识集合，其次要将该原子集合优化，取得医院数据清洗时利用的清洗序列，这也是医院数据中心数据清洗中至关重要的步骤，最后以清洗序列为基础，对医院数据中心的数据进行清洗。假设，输入为医院原始数据中心的数据，则输出为清洗过的医院数据。综上可知对医院数据中心的原始数据清洗时：

1）对医院数据中心的预处理对象进行选择。假设，要清洗的医院数据集为 $\{B_1,B_2,\cdots,B_n\}$，在知识库中选取的医院数据清洗相关知识为 $\{TQ_1,TQ_2,\cdots,TQ_n\}$，医院数据清洗属性和知识集 (B_i,TQ_i) 为相互对应的关系。由此可得知识集 TQ_i：

$$TQ_i=\{T_{i1},T_{i2},\cdots,T_{im}\} \tag{1}$$

其中：i 代表医院数据数目，m 代表数据清洗对象个数，$T_{ij}(1\leqslant j\leqslant m)$ 代表扩展树状知识库中一个结点，但不是叶结点，j 代表数据清洗对象中某一数据。对于每个 (B_i,TQ_i) 转 2)。

2）原子知识集的产生。对于 TQ_i 中的每一个知识 T_{ij}，搜寻其知识库，得到 T_{ij} 的所有原子知识：

$$T_{ij}=\{T_{ij1},T_{ij2},\cdots,T_{ijq}\} \tag{2}$$

其中：q 代表原子知识总量，$T_{ijv}(1\leqslant v\leqslant q)$ 代表原子知识。

3）对重复原子知识进行删除操作，公式为：

$$TQ_i=\{T_{i1},T_{i2},\cdots,T_{im}\}=\{T_{i11},T_{i12},\cdots,T_{i21},T_{i22},\cdots,T_{im1},T_{im2},\cdots\} \tag{3}$$

假设，F 为医院数据属性，该属性所对应的知识集可表示为 $\{T_1,T_2,T_3\}$，则其原子知识集可分别表示为：

$$T_1=(T_{11},T_{12}) \tag{4}$$

$$T_2=(T_{12},T_{21}) \tag{5}$$

$$T_3=(T_{21},T_{31}) \tag{6}$$

根据上述公式可知，数据属性 F 所对应的原子知识集表示为 $\{T_{11},T_{12},T_{21},T_{31}\}$。

4）原子知识序列的产生。对医院数据清洗的过程中，由于数据清洗知识的不同，导致数据清洗时间也就不同，因此一个数据清洗知识对应一个数据处理权重。对 3）中产生的原子知识集，将权重按照从小到大的顺序进行排列，并得到原子知识序列。假设，依据原子知识权重按照从小到大进行排序，得到序列 $T_{12},T_{31},T_{11},T_{21}$，则属性 B 所对应的原子知识序列可表示为 $(T_{12},T_{31},T_{11},T_{21})$。

5）综上所述，通过原子知识序列完成对医院数据中心数据的预处理。假设原始医疗数据库为 G，序列中原子知识个数为 H，则医院数据中心数据的预处理方法时间复杂度可表示为 $O(G\times H)$。

1.1.2 医院数据集成与数据脱敏

将 2.2 中清洗过的医疗数据，利用神经网络集成算法进行数据集成。首先对医院原始数据集 B 训练神经网络，从而得到一个分类器 N_i，将该分类器作为医院数据集成中的一个组成部分，然后利用医院原始数据集 B 的特性产生额外数据，假设得到的额外数据集为 M，通过数据集成分类器对数据集 M 进行分类，使数据集 M 中每个数据点都可以得到一个隶属于各数据类的概率分布。

假设，要对每个数据点的具体类别进行确定，则可以转换为求每个数据点所属类别的概率值，且此概率值为最大概率值。为了使后续生成的神经网络分类器以及医院数据集成中的分类器，可以有较大的差异。将数据集 M 中的每个数据点，隶属于各类的概率进行求倒数操作。假设，确定了生成数据所属类别后，与医院原始数据集 B 一起训练，成为新神经网络的医疗数据集，设置利用该数据集得到分类器 N'。

为了保障医院数据集成的准确性，将新生成的分类器 N'，加入至医院数据集成中，对数据集成分类器在医院原始数据集 B 上的分类误差率进行计算，假设该误差率小于未加入 N' 时的误差率，则 N' 可以当作医院数据集成中的组成部分，否则将 N' 丢弃。直至达到医院数据集成规模的要求，或者达到额定的迭代次数。

对医院数据集成规模以及迭代次数赋值：$i=1$，$iterations=1$，利用数据集 B 进行神经网络训练操作，则可得到分类器 N_i：

$$N_i=AA(B) \tag{7}$$

其中：A 代表神经网络训练的常数单位。将分类器 N_i 加入至医院数据集成过程中，则有 $N^*=\{N_i\}$，对数据集成分类器在医院原始数据集 B 上的误差率 λ 进行计算可得：

$$\lambda=\frac{\sum_{w_i\in B,N^*(w_i)\neq l_i}1}{m} \tag{8}$$

其中：w_i 和 l_i 分别代表医院原始数据的两个子集。利用医院原始数据集 B 的数据分布产生数据集 M，数据集 M 中数据点个数通过比例因子 ε 确定，则：

$$M=Data-Generation(\varepsilon,B) \tag{9}$$

利用得到的局部数据集成分类器，对数据集 M 进行分类，分类结果代表每个数据点，隶属各数据类别的概率分布 S，则：

$$S=Local-Ensembie-Classification(N^*,M) \tag{10}$$

其中：为了生成有差异的神经网络，依据概率分布 S，生成一个与 S 互为倒数的概率分布，即：

$$M-label=set-Class-Label(N^*,M) \tag{11}$$

将新产生的数据集 $M-label$ 和数据集 B 组合成新数据集 B，则：

$$B=B\cup M-label \tag{12}$$

利用神经网络算法对新生成的数据集 B 进行训练操作，得到的医院数据分类器为 N'，则：

$$N'=AA(B') \tag{13}$$

将式（13）得到的分类器 N' 加入到医院数据集成中，则有：

$$N^* = N^* \bigcup \{N'\} \tag{14}$$

将数据集 B 去除，产生新的数据集，$B = B - M - label$，根据新产生的数据集，对医院数据集成分类器在医院原始数据集 B 上的误差率进行计算：

$$\lambda' = \frac{\sum_{w_i \in B, N^*(w_i) \neq l_i} 1}{m} \tag{15}$$

其中：λ' 代表根据新产生的数据集，医院数据集成分类器在医院原始数据集 B 上的误差率。假设加入分类器 N' 之后，集成分类器在数据集 B 上的误差率小于未加入集成分类器 N' 后数据集成分类器的误差率，则在数据集成过程中，保留分类器 N'，否则从医院数据集成中将该分类器剔除。假设医院数据属性是连续性属性，那么对数据集 B 中的每一个连续属性均值与方差进行计算，然后根据高斯分布产生新的数据集 $Data - con$，假设医院数据属性是离散属性，则对它们的概率分布进行计算，按上述分布产生数据集 $Data - nocon$，$Data - nocon = generation(S - feature)$，将各属性数据 $Data - nocon$ 和 $Data - con$ 结合，构成新数据点 $Data$。

对上述内容进行重复操作，直至达到额定数据个数 $\varepsilon \mid B \mid$ 为止。将数据集成 N^* 中的每一个分类对医疗数据集 M 中每个数据点 w_i 进行分类，可以得到每个数据点 w_i 隶属各个类别的概率分布 $S_l^{N_l}(w_i)$，选取融合方法 E 对数据进行集成，由此可以得到数据集成分类器，对数据集 M 中的每个数据点隶属各个类的概率分布 S_l 进行计算：

$$S_l = E(S_l{}^{N_1}(w_i), S_l{}^{N_2}(w_i), \cdots, S_l{}^{N^*}(w_i)) \tag{16}$$

为了更好地保护患者隐私，保障医院的正常管理，利用 DDM 对医院数据进行脱敏操作。DDM 一般在敏感数据具有访问权限时，对数据进行脱敏，而且可以根据规划，执行对应的脱敏操作。医院数据脱敏系统主要由资源层、服务层、应用层构成，具体配置如下：

1）医院敏感数据的识别配置：对目标模型的全部数据进行智能化识别，对医院数据字段内容分析透彻，对关键词进行处理，对数据库中敏感数据进行识别。

2）数据脱敏状态监控：对医院数据脱敏系统运行状况进行监控与审计，可以及时观察到异常并且做出处理，在规定期限内将综合处理后的操作结果反馈给医院管理人员，将脱敏的需求配置尽量完善，从而提高医院数据的脱敏效率。

1.2 Paxos 算法临床全视图系统构建

1）以医院临床全视图组成结构为基础，利用扩展树状知识库对医院临床数据进行清洗。为实现 Paxos 算法临床全视图系统的构建，首先对医院临床数据来源进行统计。图 1 给出了医院临床诊疗数据的组成部分。

由图 1 可知，医院临床诊疗数据主要由：患者基本信息、医嘱信息、患者治疗过敏史、病理报告、护理记录、医学影像报告、医疗费用记录、门诊处方信息构成。医院临床数据的构成部分，对医院临床全视图的构建起到了辅助作用。

2）采用 Paxos 算法对清洗过的数据进行集成。为了获得差异比较大的神经网络，使用数据分类器对数据集合中的各个数据点分类结果，进行求倒数操作，从而获得中的各个数据点

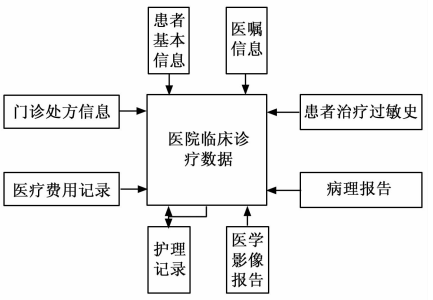

图 1　医院临床诊疗数据的组成

概率分布。当数据集成分类器中各个组成部分确定后，利用融合方法将各个分类器的数据分类结果进行融合，由此完成对医院数据的集成操作。

3）通过 DDM 对医院敏感临床数据进行脱敏操作，从而完成对医院临床全视图的构建。图 2 是医院临床全视图系统的服务层架构图。

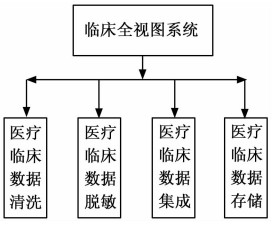

图 2　临床全视图服务层架构

分析图 2 可知，在医院临床全视图架构中，首先对医院医疗临床数据进行清洗，过滤掉冗余数据，并对清洗过的数据完成脱敏操作，然后将脱敏过的医疗临床数据进行集成和存储，最后在数据中心基础上采用 B/S 架构，将临床业务人员需要的完整详细的数据展示在临床全视图系统界面上。

2　仿真实验结果与分析

为了证明基于 Paxos 算法的医院临床全视图构建方法的实用性，需要进行一次仿真实验。在 matlab R2016b 的环境下搭建基于医院数据中心的临床全视图构建实验仿真平台。实验数据取自于我院中心机房，利用本文所提 Paxos 算法对实验数据进行全视图构建，观察其可靠性。表 1 是不同方法下数据集成时间（s）的对比。下面给出了数据集成时间（s）的计算公式：

$$\text{数据集成时间} = \frac{\text{数据集成量}}{\text{数据集成速度}} \tag{17}$$

由上述公式得出，当数据集成量为 2000 万个时，文献［8］所用时间为 10 s，本文方法所用时间为 5 s；当数据集成量为 3000 万个时，文献［8］所用时间为 17 s，本文方法所用时间为 9 s；当数据集成量为 4000 万个时，文献［8］所用时间为 26 s，本文方法所用时间为 12 s；当数据集成量为 5000 万个时，文献［8］所用时间为 32 s，本文方法所用时间为 17 s；当数据集成量为 6000 万个时，文献［8］所用时间为 41 s，本文方法所用时间为 22 s。分析表 1 可知，文献［8］所提方法进行数据集成的所用时间比本文所提 Paxos 算法所用时间多，因为文献［8］所提方法最后是以结合实例的方式，完成对医院临床全视图的构建，并未设计单独的数据集成模块，导

致数据在集成过程中没有相对应的系统控制数据集成时间，存在数据集成所用时间长的问题。而本文所提 Paxos 算法利用神经网络集成算法进行数据集成，减少了数据集成时间。

表 1 不同方法下数据集成时间对比

数据集成量/万个	文献[8]方法所用时间/s	本文方法所用时间/s
2000	10	5
3000	17	9
4000	26	12
5000	32	17
6000	41	22

分别计算文献［8］、文献［9］、文献［10］所提方法下构建的医院临床全视图，所占存储空间很大，尤其是文献［10］所提方法，全视图所占存储空间高达 400 GB，与之相比的本文方法下的医院临床全视图仅占 160 GB 的存储空间，如表 2 所示，证明了本文所提 Paxos 算法具有可靠性和稳定性。

表 2 是不同方法下医院临床全视图所占存储空间（GB）对比。

表 2 不同方法下医院临床全视图所占存储空间对比

全视图构建方法	全视图所占存储空间/GB
本体论	280
XML 中间件技术	320
Datalog 规则	400
Paxos	160

图 3 是不同方法下医院数据清洗效率（%）对比。下面给出了数据清洗效率（%）计算公式：

$$\text{数据清洗效率} = \frac{\text{总数据量}}{\text{数据清洗时间}} \times 100\% \tag{18}$$

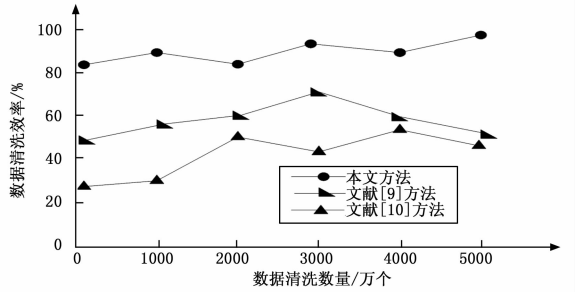

图 3 不同方法下数据清洗效率对比

从图 3 可以看出，数据清洗效率随着数据清洗量的不断增加而变化，文献［9］和文献［10］所提方法的数据清洗效率曲线波动很大，而且数据清洗效率很低，本文所提方法的数据清洗效率高且效率曲线起伏不大，而本文 Paxos 算法数据清洗效率较高，这主要是因为在利用 Paxos 算法进行数据清洗时，采用了扩展树状知识库完成对医院数据的清洗，使得 Paxos 算法具有较高的清洗效率，进一步证明了本文所提方法的整体有效性。图 4 是不同方法下数据脱敏覆盖率（%）的对比。下面给出了数据脱敏覆盖率（%）计算公式：

$$\text{数据脱敏覆盖率} = \frac{\text{已脱敏数据量}}{\text{应脱敏数据量}} \times 100\% \tag{19}$$

脱敏数据对医院数据的管理非常重要，分析图 4 可知，文

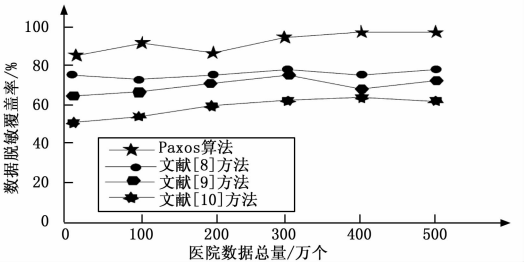

图 4 不同方法下数据脱敏覆盖率对比

献所提方法的脱敏覆盖率相对集中，表示它们的脱敏性能大致相同，覆盖率基本维持在 78%以下，而本文所提 Paxos 算法的脱敏覆盖率几乎在 82%以上，本文最低脱敏覆盖率与文献最高脱敏覆盖率相差 4%，证明了本文方法对医院敏感信息的保护相对稳定可靠。

仿真实验证明，Paxos 算法可以准确地对基于医院数据中心的临床全视图进行构建。

3 结束语

采用传统方法对基于医院数据中心的临床全视图系统进行构建时，无法构建出准确详细的全视图，存在医护人员对医疗数据查询时，查询结果不理想，无法在同一界面看到所有相关信息数据的问题。提出一种基于 Paxos 算法的医院数据中心临床全视图构建方法。并通过仿真实验证明，Paxos 算法可以准确地对医院临床全视图进行构建，具有优秀的应用价值。后期我院将与上海柯林布瑞信息技术有限公司合作，利用 Paxos 算法将基于临床数据中心的临床全视图系统实施落地，更好的服务于临床。

参考文献：

[1] 奈存剑，任宇飞，李 金，等. 医院临床数据中心建设与应用［J］. 中国医院管理，2014，34（5）：53－54.

[2] 张文捷，蒋 抒，张 民. 我院绿色数据中心建设的实践［J］. 中华医院管理杂志，2016，32（5）：394－396.

[3] 高 明，唐 顺，徐福文. 医院数据挖掘平台中 X－11－ARIMA 预测模型的应用研究［J］. 中国卫生统计，2016，33（1）：139－141.

[4] 于国泳，杨 薇，谢雁鸣，等. 医院信息系统数据库 72772 例 2 型糖尿病患者临床特征分析［J］. 北京中医药大学学报，2014，37（12）：851－857.

[5] 吴正一，崔迎慧，陆 耀，等. 以临床数据仓库为核心的医院大数据平台构建［J］. 中国医院管理，2015，35（11）：13－15.

[6] 杨 莘，韩斌如，应 波，等. 基于信息数据中心决策支持平台构建护理质量评价体系［J］. 中华护理杂志，2015，50（1）：10－13.

[7] 陈 川. 基于学科元数据中心的知识服务平台建设研究与应用［J］. 情报理论与实践，2014，37（5）：57－60.

[8] 侯佳音，史淳樵. 云计算技术在医院的信息化建设中的应用研究［J］. 电子设计工程，2016，24（5）：35－39.

[9] 刘 伟，赵一林. 数据包络分析在地市级中心医院综合效率分析中的应用［J］. 中国卫生统计，2014，31（5）：896－898.

[10] 宗酉明，詹伟国，毕鲁佳，等. 医院影像系统图文大数据云存储的实践应用［J］. 中华医院管理杂志，2015，31（12）：940－942.