·卫生管理· doi:10.3969/j.issn.1671-8348.2019.14.041

医疗大数据平台研究与实践*

龚 军[1],孙 喆[2],向天雨[3],王惠来[3△]

(1.重庆医科大学医学数据研究院 401331;2.医渡云(北京)技术有限公司,北京 100191;
3.重庆医科大学附属大学城医院信息中心 401331)

[摘要] 研究医疗大数据平台搭建方法,以及与和医学相关的技术和实践意义。从医疗大数据平台的基础架构、多源异构数据的统一及标准化、智能患者索引技术、数据质量监控与评价技术、医疗大数据平台的分析技术等方面进行分析。同时介绍重庆医科大学区域医疗大数据平台建设概况和在医院管理和科研上的应用状况。重庆医科大学医疗大数据平台汇集重庆市多家大型三甲医院业务数据,目前已成功上线且运行稳定,基于此数据平台开展了多项数据应用。医疗大数据平台在促进医院互联互通,提升医院医疗质量和效率有积极促进作用。

[关键词] 大数据;医疗健康;数据集合;数据挖掘

[中图法分类号] R-05 [文献标识码] B [文章编号] 1671-8348(2019)14-2504-04

随着信息时代的到来,互联网技术、存储技术、信息技术等在人类生产、生活中大规模的应用。各行各业采用的信息系统中产生了海量的数据,并且仍在急速的增长。据统计,2006 年全球共新产生约 180 EB 数据,2011 达到了 1.8 ZB。而据互联网数据中心(IDC)预测,到 2020 年,全世界数据总量将增长 44 倍,达到 40 ZB[1]。大数据时代的到来将带给各行业新一轮的变革,对各行业的发展将是一个新的契机。由于医疗行业的特殊性,医疗机构中保存着大量电子病历数据和电子健康数据。但是,由于有关数据利用的规范还不够标准,相关技术尚欠缺。此外,海量的医疗数据仍储存在各医院的数据库中且各自互不联通,数据的探索和利用十分困难。区域医疗大数据平台的建设能够解决信息孤岛的问题,使区域内的数据互联互通。基于医疗大数据能够对真实世界进行研究,并从数据中挖掘出有价值的信息。

1 区域医疗大数据平台的架构

区域医疗大数据平台基于 Hadoop 环境框架搭建,在大量原始数据的基础上对数据进行汇集、整合、清洗、计算及分析与应用,形成大数据平台架构。

1.1 医疗大数据平台基础架构 医疗大数据平台基础架构大致可以分为数据采集、数据存储、数据处理、数据分析、数据应用及系统控制。基本思路:大数据平台汇集医院的业务数据进行分布式存储,对业务数据进行脱敏、清洗、结构化、归一和质控等操作,以数据需求为维度存入分布式数据库,并在其基础上构建数据分析平台和管理应用平台。医疗的数据平台基础架构见图 1,从下到上为系统架构线路,各部分为平台搭建相关系统名称。

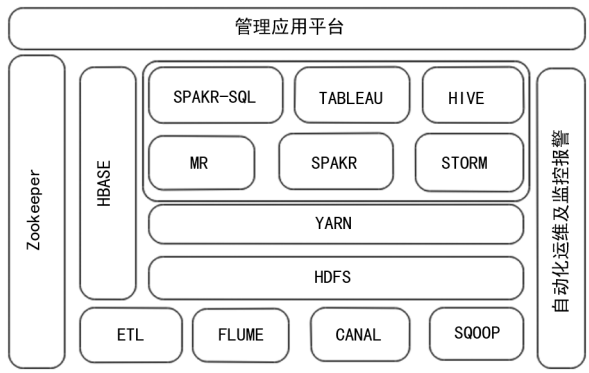


TABLEAU:一种专业数据可视化软件

图 1 医疗大数据平台基础架构图

数据采集由 ETL、FLUME、CANAL 等技术实现。ETL 批量采集医院离线数据,FLUME、CANAL 采集医院的业务系统实时数据。HDFS、HBASE 技术相互协同实现数据的存储功能。在资源管理器 YARN 和 AZKABAN 的协调下,MR、SPAKR、STORM 技术分别对大批量数据和实时数据进行分布式计算。管理应用平台通过 R、SPAKR-SQL、数据挖掘等手段对数据进行分析利用。

1.2 大数据平台核心技术 大数据平台核心技术围绕着怎样处理医院的大规模数据展开,由于医疗行业的特殊性,区域医疗大数据平台的建设将应用传统的大数据技术并在此基础上进行创新。

1.2.1 多源异构数据的统一及标准化 目前全国共有上百家 HIS 系统厂商,由于各厂商运用的技术架构和处理的问题不同,其系统中数据标准和结构也多种多样。大数据平台需要采集 HIS 框架中的全量数据。因此,对多源异构数据的采集、汇总、结构化是建立医

* 基金项目:重庆市科委重大主题专项(cstc2015shms-ztzx10011)。 作者简介:龚军(1996-),在读硕士,主要从事医疗大数据方面的研究。 △ 通信作者,E-mail:804225405@qq.com。

疗大数据平台亟待解决的问题。区域医疗大数据平台采用多源异构数据统一及标准化技术处理这些“结构复杂”的数据。

医疗大数据平台根据数据的标准化程度和用途不用分为原始数据层、结构化层、应用层。原始数据层是 HIS 系统直接写入到 Hadoop 集群的数据，结构化层是原始数据经过结构化、标准化的数据，应用层是大数据平台根据某项业务需求按不同纬度抽取结构化层数据而得到的集成数据。多源异构数据的统一及标准化技术作用于原始数据层和结构化层之间。多源异构数据的处理框架见图 2。

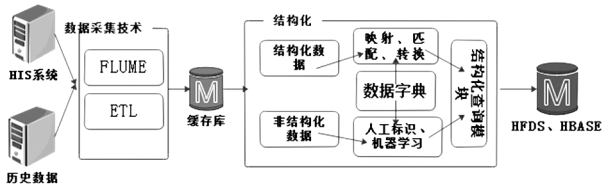


图 2　多源异构数据处理框架

多源异构数据的采集是运用 LSA 算法和 redo-log 分析对任意厂商提供的数据库进行内容识别。利用在医院搭建的前置机将医院的电子病历等数据写入到 Hadoop 的分布式文件系统(HDFS)中。针对于医院产生的实时数据，利用 FLUME 系统将数据写入到 Hadoop 集群。把医院多个系统产生的多种数据归一成以下全面覆盖医院各系统、各场景的可扩展的数据模型，并对其进行结构化和归一。医院信息系统数据汇集图见图 3，模型重构下面代表医院 HIS 系统及各个子系统，模型重构上面代表数据的分类及存储形式。

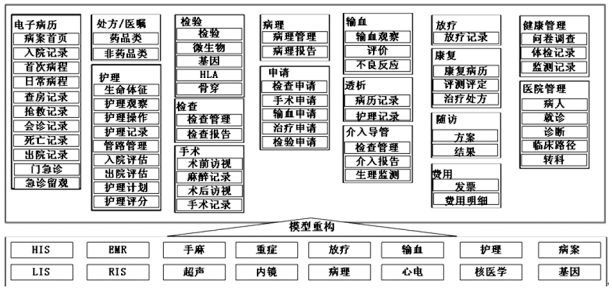


图 3　HIS 系统数据汇集框架图

数据按其结构化程度分为结构化数据、非结构化数据和半结构化数据。数据处理以医院为单位，即对每个医院都建立一套数据抽取的标准。这套标准是专业的 ETL 人员结合数据字典对该医院每个字段建立相应的提取规则，并经过多次质控而形成的。数据字典来源于国内外通用的数据标准，如 HL7、CDA、卫计委电子病历基本架构与数据标准、ICD-9/10 等，也有国内外专家经研究共同决定的标准。

对于结构化的数据，主要处理“标准化”的问题。如出院记录中的出院日期，医生可能有多种写法，在数据处理阶段需要将其映射为标准的日期时间格式。又如出院诊断的胃癌分型，数据字典中结合国际标准与权威专家的评估把胃癌分为 5 个类型，从而把临床医生不标准诊断名称全部标准化。

非结构化数据的处理首先是使其“结构化”。非结构化数据是由自然语言构成的文本，如手术经过或一诉五史。非结构化数据使其结构化采用的技术主要是人工标识加机器学习。从 Hadoop 集群中采样部分数据作为训练集，利用人工标识技术将需要在自然语言中获取的字段标出，机器学习算法对训练集作分词、词性标注、命名实体识别、依存句法分析处理，建立非结构化数据结构化处理模型，从而应用于非结构化数据的处理。对已经结构化的非结构化数据同样可以进行归一和标准化。其中，人工标识技术是按照电子病历结构拆分成节点，对其重要的信息人工进行标注并以 KV 键值形式展现。非结构化数据处理流程见图 4。

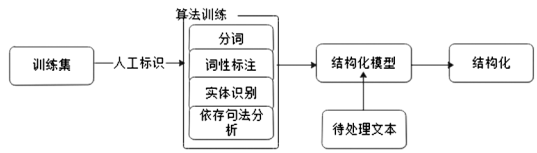


图 4　非结构化语言处理流程

1.2.2　数据质量监控与评价技术　医疗大数据平台汇集医院信息系统的全量数据，涵盖各个子系统，数据量多且复杂。由于医院的数据产生于临床医生及护士的操作，在数据科学方面缺少专业知识及系统自身的问题等一系列因素，导致了大数据平台的数据会存在空值、违规、错误等问题。多源异构数据统一及结构化技术能解决数据的汇集和标准的问题，但无法解决数据的完整性、一致性、准确性、稳定性等问题。

数据质量监控技术的基本思路是通过数据质量评价标准和定量定性的分析手段找出数据存在的质量问题，指导数据优化的实施。数据质量控制的基本框架见图 5。

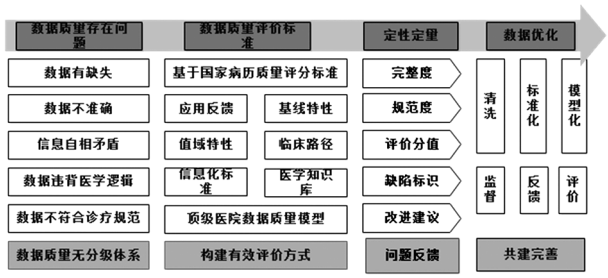


图 5　数据质量评价技术框架

数据质量监控技术作用于结构化层或者应用层。根据质量评价标准事先设立质量控制指标和每个指标的阈值，质控结果一旦超过了阈值则会针对该诊断进行报警，具体质控指标见图 6。针对每个质控指标建立对应的算法，例如：空值率＝(字段缺失数＋空值数量)/应有数据量×100%。根据指标和算法编写 SQL 语言，提取 HDFS 中结构化全量数据和应用层数据，分析结构化字段是否有数据缺失、数据不准确、

信息自相矛盾等问题，汇总生成数据质量评估报告，反馈给数据 ETL 人员。ETL 部门根据此报告调整数据提取、清洗算法，再次提交质控，从而使数据的合格率超过设定的阈值。

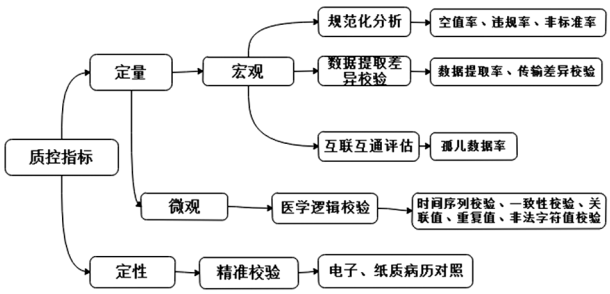
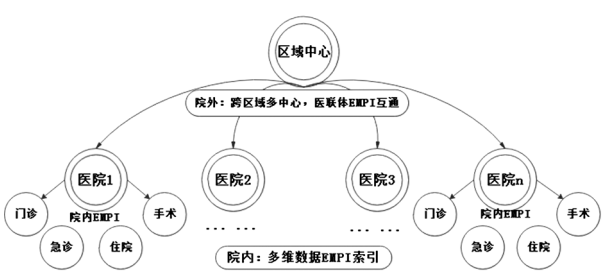


图 6 数据质量控制指标

1.2.3 智能患者索引技术 目前的 HIS 系统中，医院通常把患者挂号时产生的门诊号或者住院号当做患者主索引。由于目前每个区域都有多家不同性质的医疗服务机构，如综合医院、社区卫生服务中心、疾控中心、专科医院等，各个医院采用不同的 HIS 系统且都互不联通，则产生的患者主索引也会有很大的差异。因此，在建立医疗大数据平台时，无法根据现有的技术对数据以患者纬度进行统一。

区域医疗大数据平台在利用 Hadoop 技术实现区域内患者数据互联的基础上，对医院的全量数据进行组织梳理，提取完整的患者信息，根据数据间的业务关系，以患者为维度将区域内患者的数据形成一个整体，从而形成患者完整的就诊档案，通过主索引可以检索出患者在各个地方的就诊记录。

智能患者主索引技术能让医院内部仍旧使用住院和挂号时自动生成的住院号和门诊号作为患者就诊的唯一标识，在各医院之间使用社保卡和身份证号码作为识别患者身份的唯一标识。门诊号和就诊卡号能够保证医院作为医疗数据平台的一个数据节点高效的运行，在互联互通的基础上保持自己的独立性。区域医疗大数据平台的控制中心提取出各个医院 HIS 系统的患者身份信息（身份证号、社保卡号、地址等）建立患者索引，关联患者在医院的就诊数据。其中，提取患者多种身份信息是为了保证无论患者采用什么方式挂号就诊，系统都能采集到该患者的就诊数据。智能患者索引架构图见图 7。

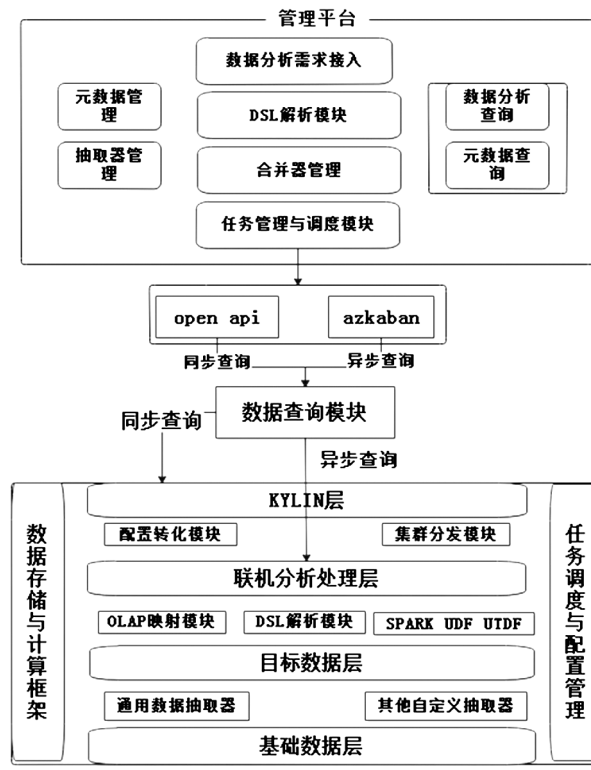


EMPI：患者主索引

图 7 智能患者索引技术图

1.2.4 大数据的分析技术 大数据分析技术是区域医疗大数据平台建设的关键技术，是挖掘隐藏在海量数据中的知识达到数据价值化的重要手段。大数据分析技术基本思路：医院、科研工作平台及数据质控平台产生数据查询和分析需求，使用数据分析引擎提取目标数据，基于需求分析数据。

数据应用平台与数据管理平台和数据存储框架之间搭建应用程序接口，将数据提取指令传递给管理平台，经分析后制订数据提取任务，以同步和异步的提取方式提取目标数据。数据迁移路线：数据抽取器抽取 HDFS 数据为目标数据层，目标数据层由 OLAP 映射模块转化为 OLAP 数据层，OLAP 数据层通过预计数等方式转化为 KYLIN 数据立方体存入 HBASE。OLAP 数据层数据支持通过 Spark 进行离线异步查询，KYLIN 层数据支持在线同步查询。数据分析技术支持除常规的统计分析外，还可以为数据挖掘、模式识别、机器学习、并行处理等先进大数据应用提供数据支持和平台支持，从而得到更深层次的知识和领域信息。大数据平台中数据分析的框架见图 8。



各部分：数据分析平台的组成模块

图 8 医疗大数据分析平台整体架构

2 区域医疗大数据平台实践

2.1 重庆医科大学医疗大数据平台建设概况 截至 2018 年 9 月，重庆医科大学与医渡云（北京）技术有限公司合作建立的重庆医科大学医疗大数据平台已经成功上线，共汇集了重庆医科大学附属大学城医院、重庆医科大学附属儿童医院、重庆医科大学附属第二医院、重庆医科大学附属永川医院等多家医院数据。

涵盖 18 049 608 例患者就诊数据，36 109 970 人次的就诊数据，时间跨度为 1999－2018 年。重庆医科大学医疗大数据平台汇集其附属医院业务数据，开发 ETL 数据平台对所有数据进行处理和质控，通过查找国内外权威的医学知识和政策法规建立知识库和标准库。基于分布式数据库建立索引库，提升数据查询效率。在分布式数据库的基础上建立数据应用平台，探索数据的应用。重庆医科大学医疗大数据平台已经运用到医院管理、绩效评价、医学科研等多方面，其中，基于大数据平台建立的重庆医科大学科研数据平台也已投入使用，基于此数据已发表高质量科研论文多篇。重庆医科大学医疗大数据平台架构图见图 9。

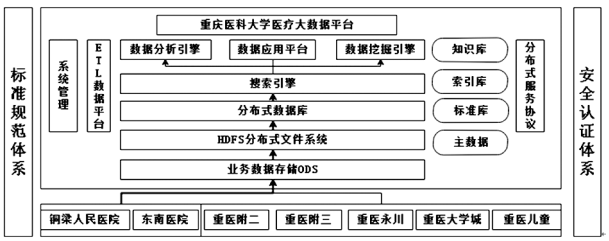


图 9　重庆医科大学医疗大数据平台架构图

2.2　医院管理　医疗大数据平台汇集某地区内大部分医院的诊疗数据，由控制平台全面管理、协调和控制。原卫计委等相关部门根据大数据平台汇集的数据，通过各项指标宏观地掌握该地区的医疗情况，据此有针对性地制订政策进行调控。各医院根据大数据平台处理后的数据进行可视化分析，改进医院的医疗服务质量。

目前基于大数据平台开发的医院管理平台在重庆医科大学附属医院已成功上线并进行了应用，涵盖医院运行管理、DRGs 绩效管理、医院等级评审、医院机构对比等功能。大数据平台采集电子病历中病案首页和费用信息，以 DRGs 相关指标对数据进行筛选并进行标准化，建立 DRGs 的数据集市，通过数据集市对费用中的住院总费用、检查检验费、手术费、床位费等和出院诊断中的主要诊断、次要诊断、并发症、手术操作等进行关联性分析，从而建立 DRGs 费用库。医院管理人员就可以基于 DRGs 费用库对各科室进行绩效评价，科室也可以据此对医生进行绩效评价。目前，基于大数据的 DRG 绩效评价正在重庆医科大学附属大学城医院初步应用。

2.3　科研应用　目前，重庆医科大学基于医疗大数据的科研平台已成功上线，具有“数据概览”“病历搜索”“知识全库”“我的科研”等功能。数据概览能对整个大数据平台的数据进行全局的展示。病历搜索对 35 721 967 份病历进行搜索，以患者或病历为维度进行展示。“我的科研”功能支持创建科研项目，按项目要求设置筛选条件，并支持导出数据。知识全库主要汇集国内外相关的医学文献，以中文核心期刊和 Pubmed 为主。重庆医科大学目前正与国家不良反应监测中心开展药物不良反应主动监测的项目，基于此平台的数据在药源性肝损伤和单种药的不良反应研究上已经取得成果。

3　小　　结

随着大数据时代的到来及信息技术在医疗行业的应用，各个医院积累了大量的数据，但仅针对一家医院的数据进行研究不能称之为大数据，医疗行业需整合区域所有的医疗信息，以更多的数据反映更加真实的医疗情况。大数据平台只是进行数据分析与应用的工具，要想充分利用这些数据，需要对数据进行分析、挖掘和深层次的算法研究和利用人工智能技术，才能为医疗行业提供有价值的决策和科学研究，充分发挥数据的价值。最终让医疗行业向效率高、技术强、费用低的方向发展。

参考文献

[1] 光环大数据. 大数据分析 S 大数据下网络视频类用户行为分析[EB/OL]. [2018-07-19]. https://zhidao.baidu.com/question/552507231674314972.html.

[2] 代涛. 健康医疗大数据发展应用的思考[J]. 医学信息学杂志，2016，37(2)：1-8.

[3] 周雪晴，罗亚玲. 信息化建设中医疗大数据现状[J]. 中华医学图书情报杂志，2015，24(11)：48-51.

[4] 沈韬，崔泳. 医疗大数据：期望与现实[J]. 中国数字医学，2015，10(7)：2-4，32.

[5] 汪鹏，吴昊，罗阳，等. 医疗大数据应用需求分析与平台建设构想[J]. 中国医院管理，2015，35(6)：40-42.

[6] 罗旭，刘友江. 医疗大数据研究现状及其临床应用[J]. 医学信息学杂志，2015，36(5)：10-14.

[7] 张振，周毅，杜守洪，等. 医疗大数据及其面临的机遇与挑战[J]. 医学信息学杂志，2014，35(6)：1-8.

[8] 邹北骥. 大数据分析及其在医疗领域中的应用[J]. 计算机教育，2014(7)：24-29.

[9] 蔡佳慧，张涛，宗文红. 医疗大数据面临的挑战及思考[J]. 中国卫生信息管理杂志，2013，10(4)：292-295.

[10] 王红迁，汪鹏，王飞，等. 基于 Hadoop 架构的医疗大数据平台应用实践和思考[J]. 医学信息学杂志，2017，38(9)：27-31.

[11] 乐颖，刘南. 医疗大数据平台的建设路径[J]. 电子技术与软件工程，2018(3)：198.

[12] 张伟. 医疗大数据平台数据高并发方案设计与关键技术分析[J]. 信息技术与网络安全，2018，37(4)：18-22.

[13] 陈素琼，王惠来，向天雨. 医疗大数据应用现状研究[J]. 中国数字医学，2017，12(9)：30-31，55.

（收稿日期：2019-02-24　修回日期：2019-05-12）