基于 Spark 的疾病诊疗自助服务平台设计*

陈桂宏 蔡庆玲

（广东东软学院计算机系 佛山 528225） （中山大学工学院 广州 510006）

〔摘要〕 针对当前医疗资源短缺的问题，利用云计算、大数据等技术，构建基于 Spark 的疾病诊疗自助服务平台，介绍平台总体架构、功能模块并进行测试，实现疾病诊断、治疗方案推荐以及疾病预测。

〔关键词〕 云计算；医疗大数据；疾病诊疗；隐私保护

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.08.007

Design of Self－service Platform for Disease Diagnosis Based on Spark *CHEN Gui－hong, Department of Computer Science and Technology, Neusoft Institute Guangdong, Foshan 528225, China; CAI Qing－ling, School of Engineering, Sun Yat－sen University, Guangzhou 510006, China*

〔**Abstract**〕 The paper aims at the current problem of medical resource shortage, builds the self－service platform for disease diagnosis based on Spark by taking advantage of cloud computing, big data and other technologies, introduces the overall architecture and functional modules of the platform and tests the platform, in order to achieve disease diagnosis, recommendation of therapeutic schemes and disease prediction.

〔**Keywords**〕 Cloud computing; Medical big data; Disease diagnosis; Privacy preserving

1 引言

近年来，随着我国人口老龄化日益加剧及“亚健康”[1]人群日益增多，医院所承载的责任越来越大。医疗行业面临着严重的医疗资源短缺问题，医院患者也随之越来越多，从而导致看病难，挂号难。改变传统的医疗卫生服务模式，利用现代信息技术手段，搭建医疗服务平台，使得各类人群都能自助地获得远程医疗、慢性病监控和疾病预测等医疗服务，是缓解当前医疗资源缺乏的有效途径。移动互联网、物联网、可穿戴设备以及医疗信息化的发展，个人医疗和健康数据呈现几何级增长，这些海量、异构、实时、来源多样化的医疗大数据，一方面为自助医疗所需的数据分析提供了基础，而另一方面数据被封闭在各个医院或机构，也限制了医疗大数据的应用价值。云计算是大数据成长的驱动力，其强大的计算和存储能力[2]为医疗服务平台的建设提供了必要条件，但同时也带来存储数据的安全问题。而且通过大数据分析技术实现健康状况评估和疾病预测也会带来严重的隐私问题[3]。

目前，医疗大数据的应用已经引起了广泛关注，出现一些大数据医疗应用平台。如微软的个人健康管理平台（Health Vault）[4]，用户可以便捷地

〔修回日期〕 2017－05－25

〔作者简介〕 陈桂宏，博士研究生，讲师，发表论文多篇。

〔基金项目〕 广东省协同创新与平台环境建设项目（项目编号：508300984106）；广东省青年创新人才类项目（项目编号：2016KQNCX192）；佛山市科技创新项目（项目编号：2016AG100382）。

通过网页上传自己的健康数据并进行存储、检索和分享；Lin 等[5]提出的疾病家庭诊断系统，用户可以方便地在家中进行疾病诊断，能够为不同的人群，特别是老年人和慢性病患者提供疾病预防的知识和方法。在考虑用户隐私方面，Xu 等[6]提出基于云计算的远程医疗监控系统，利用云平台实现数据存储、疾病诊断，通过租户机制保障个人健康数据的安全和隐私；Mathew 等[7]构建了能够保证用户隐私的疾病决策树，用于疾病诊断，但仅能保证训练数据集的隐私；Liu 等[8]采用朴素贝叶斯分类法实现了在线疾病诊断，既能保护用户的查询信息又能保护服务提供方的诊断模型；Li 等[9]、Zhou 等[10]利用基于属性的加密方法实现了云环境下个人健康数据的隐私保护。由此可见，在医疗大数据的知识发现和隐私保护方面，已经有一定的研究成果，但众多的研究仍然处于实验阶段，在实用性、效率、隐私性等方面还存在很大挑战。鉴于此，本研究构建保护用户隐私的疾病诊疗自助服务平台，利用云计算框架对医疗大数据进行存储、查询、分析和挖掘，在保护用户隐私的同时，实现疾病诊断、治疗方案推荐以及疾病预测。

2 平台设计

本研究基于现有平台的技术优势以及所存在的挑战，构建疾病诊疗自助服务平台，实现疾病诊断、预测以及治疗方案推荐，旨在解决以下几方面问题：（1）实用性。平台设计既要考虑功能，又要考虑用户的知识背景，使得平台能面向所有人群，解决医疗问题。（2）效率。选择合适的大数据处理工具，设计大数据并行处理算法，是平台实现的关键。（3）可伸缩性。医疗大数据实时变化，呈指数级增长，平台的计算能力必须自适应地扩展；另外随着处理需求的增多，相应的算法也应能方便地集成到平台中。（4）隐私性。隐私是大数据应用所关注的热点问题，虽然目前在隐私保护方面已有一定的研究成果，但在还存在很大的挑战，解决用户的隐私保护问题，是平台得以推广的关键。

3 平台实现

3.1 总体架构

本平台核心模块由隐私保护、存储、分析几部分构成，见图 1。核心数据处理、计算工作在云端完成，云端构建有 3 个集群，分别对应数据隐私保护、存储、分析功能。首先，来自医院的电子病历、个人的电子健康记录以及相关研究机构的数据，如药物研究数据等，传输到数据预处理集群 1，经过隐私处理模块，进行匿名和加密，以保证数据隐私安全；经过隐私处理后的数据传入集群 2，将数据统一为 Lucene 文档格式并建立文档索引，存储在 HDFS 分布式文件系统中；终端数据用户通过身份认证后，对集群 3 发出数据分析请求，如患者、医生可通过症状描述查询病症；根据某种疾病的治疗路径和效果，得到最优的治疗方案推荐。

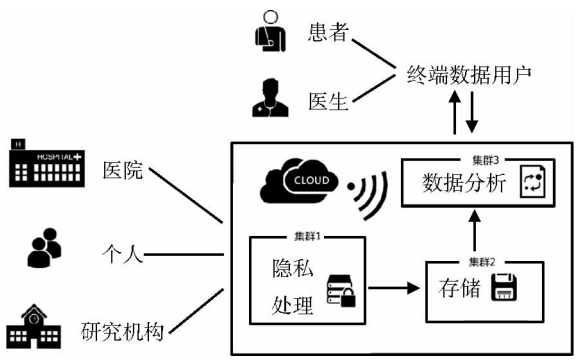

图 1 诊疗平台总体架构

3.2 模块

3.2.1 数据隐私保护 （1）计算平台。见图 2，为实现自助诊疗平台所涉及的批处理、实时处理和迭代计算，本平台选择 Spark 作为云计算框架。Spark[12]是新一代快速通用的大数据计算引擎，其基于内存计算的特点，将计算的中间结果存储在内存中以减少磁盘 I/O 开销，非常适合迭代计算、实时处理和交互式处理。另外，与 Spark 相结合，数据存储基于 HDFS 分布式文件系统。（2）隐私处理。隐私处理不仅需要考虑隐私保护程度，还要考虑数据可用性，而匿名[11]是二者折中的有效方法。

本平台基于 Spark 框架，以 k－匿名为基本模型，集成多种隐私保护匿名算法，算法选择由隐私管理模块根据用户对隐私保护的要求自主控制。由于同一源数据集可能被匿名为不同数据集，这些数据集联合之后也存在隐私泄露风险，因此采用 Zhang 等[13]中加密部分数据集的方法，在尽量减少计算开销的同时加强隐私保护。

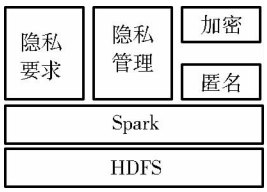

图 2 数据隐私保护模块

3.2.2 *数据存储* 本平台采用 Lucene[14]作为搜索模型。Lucene 是 Apache 的开源全文检索引擎工具包，其核心组成为索引和搜索，利用 Lucene 提供的应用程序接口将异构医疗大数据映射为 filed－value 格式的文档，存储在 HDFS 中，为建立索引做好准备。参考 Lin 等[5]建立索引的方法，本平台利用 Spark 进行分布式索引建立。索引建立与搜索节点有关，假设搜索节点个数有 $N \times M$ 个，见图 3，N 表示索引划分数，M 表示并发搜索数。首先，采用 parallelize（）函数将 Lucene 文档分为 P 份，P 值大小与集群节点个数有关；划分之后的文档送到各个工作节点，工作节点在 Map 阶段为收到的文档建立索引，将索引分成 N 份，在 Reduce 阶段将具有相同 key 值的索引整合在一起。至此，全局数据的索引以 N 个片段的形式存储在 HDFS 中，支持并行搜索，提高在线数据分析速度。

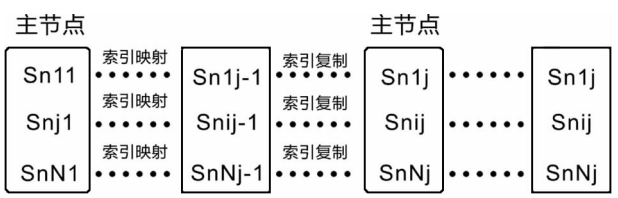

图 3 搜索节点阵列

与 Lin 等[5]不同，本平台列节点索引采用映射加复制相结合的机制，减少占用存储空间。第 1 列 N 个节点作为主节点对应复制索引文件的 N 个片段，之后 $j-1$ 列所操作的索引文件均为第 1 列的映射；当第 1 列节点的 IO 访问达到限制数时，触发复制索引文件到第 j 列，之后第 j 列又再作为后续列节点的主节点。

3.2.3 *数据分析* 数据分析模块由离线计算模块和在线计算模块两部分构成。离线计算模块从 HDFS 历史数据中建立疾病诊断和预测模型，集成机器学习、统计分析、推荐算法；在线计算模块一方面根据用户症状描述搜索相似患者，提供实时病症诊断，另一方面根据各种实时检测数据，如血压、心率等，预测疾病。

4 平台测试

用户交互界面设计，见图 4，用户通过 Web 与云平台进行交互。注册用户可保存自己的个人信息，通过输入症状得到疾病诊断结果。历史数据包含用户的查询记录以及身体各项指标的检测数据，用于评估用户身体状况，在疾病预测栏显示预测结果。图 4 中在疾病诊断栏，列有人体主要部位可能出现的症状，选择相应症状，确认提交后可得到疾病诊断结果以及推荐的治疗方案；如果所列症状不足以表述用户身体状况时，用户可以在其他栏对自身症状进行描述。

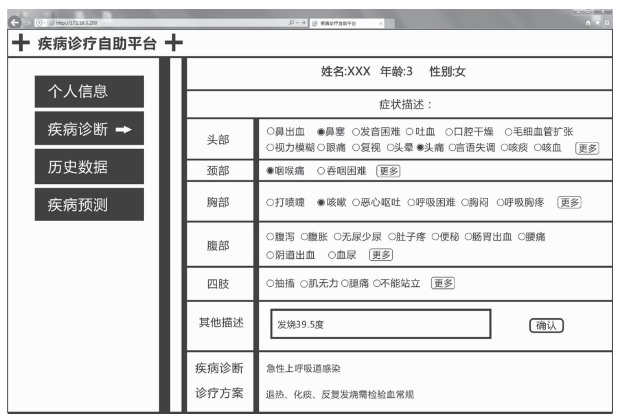

图 4 用户交互界面

5 结语

本研究采用开源大数据处理架构，构建了疾病

诊疗自助服务平台，具有以下特点：（1）实用性。平台具有疾病诊断、预测以及治疗方案推荐功能，符合移动医疗需求；平台通过 Web 端与云平台进行交互，界面设计能够很好地引导用户尽可能准确地描述症状，而不需要过多医学知识背景。（2）高效。利用云计算框架实现匿名、加密，提高计算速度，通过构建基于 Lucene 的分布式搜索集群提高搜索速度。（3）可伸缩性。一方面通过扩展节点增加了平台的存储及计算能力；另一方面提供算法接口，方便集成各种数据分析算法，扩展平台的医疗应用范围。（4）隐私性。采用匿名和加密相结合的方法，在保证数据可用性的同时，增加数据隐私安全；通过加密部分数据集的方法，减少计算开销。在下一步工作中，将基于本平台，进一步优化系统架构，扩展隐私保护、数据分析和挖掘算法，以提高系统功能和性能。

参考文献

1 Ding HJ, He JC, Wang WW, The Sub - health Evaluation Based on the Modern Diagnostic Technique of Traditional Chinese Medicine, Education Technology and Computer Science [C]. First International Workshop on ETCS, 2009: 269 - 273.

2 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51 (1): 107 - 113.

3 冯登国，张敏，李昊. 大数据安全与隐私保护 [J]. 计算机学报，2014，37 (1)：246 - 258.

4 Microsoft HealthVault [EB/OL]. [2016 - 09 - 10]. http://en.wikipedia.org/wiki/Microsoft_HealthVault.

5 Lin WM, Dou WC, Zhou ZJ, et al, A Cloud - based Framework for Home - diagnosis Service Over Big Medical Data [J]. Journal of Systems and Software, 2015, 102 (C): 192 - 206.

6 Xu BY, Xu LD, Cai HM, The Design of an m - Health Monitoring System Based on a Cloud Computing Platform [J]. Enterprise Information Systems, 2015, 11 (1): 1 - 20.

7 Mathew G, Obradovic Z, A Privacy - preserving Framework for Distributed Clinical Decision Support [C]. ICCABS'11 Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences, 2011: 129 - 134.

8 Liu XM, Lu RX, Ma JF, et al. Privacy - preserving Patient - centric Clinical Decision Support System on Naive Bayesian Classification [J]. IEEE Journal of Biomedical & Health Informatics, 2016, 20 (2): 655 - 668.

9 Li M, Yu SC, Zheng Y, et al. Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute - Based Encryption [J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24 (1): 131 - 143.

10 Zhou J, Lin XD, Dong XL, et al, PSMPA: patient self - controllable and multi - level privacy - preserving cooperative authentication in distributed m - healthcare cloud computing system [J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 26 (6): 1693 - 1703.

11 Chen GH, Cai QL, Zhan YJ, Approaches on Personal Data Privacy Preserving in Cloud: a survey [C]. BIGDATA 2016, 2016: 36 - 43.

12 Zaharia M, Chowdhury M, Franklin MJ, et al. Spark: cluster computing with working sets [C]. Hot Cloud'10 Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 2010: 10.

13 Zhang XY, Liu C, Nepal S, et al. A Privacy Leakage Upper Bound Constraint - based Approach for Cost - effective Privacy Preserving of Intermediate Data Sets in Cloud [J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24 (6): 1192 - 1202.

14 Hatcher E, Gospodnetic O, McCandless M. Lucene in Action [M]. New York: Manning Publications, 2010.