

大数据背景下核能安全人因数据采集与处理系统框架

王以群,吴 博^①,方小勇^②

(南华大学 人因研究所,湖南 衡阳 421001)

[摘 要] 核能人因安全数据分析在核安全管理中的地位越来越重要。然而,不同人因分析领域、目标和方法,有不同人因数据需求,同时在大数据背景下,人因数据分析面临着海量、低价值密度人因数据的集成与抽取;非结构化数据的处理与融合等问题。为解决这些问题,应用大数据技术,根据人因数据采集与处理的特点,建立了核能安全人因大数据采集、存储、分析和处理系统框架。该系统框架由数据源层、数据预处理、人因大数据处理中心、人机交互系统等部分组成。

[关键词] 核能安全; 人因数据; 数据采集与处理; 大数据技术

[中图分类号] TM623;TP311.13 **[文献标识码]** A **[文章编号]** 1673-0755(2018)01-0012-05

大数据泛指大规模、超大规模的数据集,因可从中挖掘出有价值的信息而倍受关注,但传统方法无法进行有效分析和处理。一般而言,大家比较认可大数据从早期的3V、4V说法到现在的5V(新增了Value)。大数据的5个“V”,业界将其归纳为Volume, Variety, Value, Velocity, Veracity^[1]。而前三个(海量、异构、低价值密度)是对大数据的属性特征的描述,后两个是对大数据的需求特征的描述。也就是说,理论上只要满足前三个特征的数据都可以采用大数据进行处理,但是还要同时满足高速处理及准确性(Velocity和Veracity)的数据处理需求。

核能安全人因研究是核电安全的重要内容,通过人因分析不仅能防范人因失误还能发现核能安全需求^[2],对核能安全有重大意义。可是人因数据本身不仅包括核电厂的运行数据、环境数据、工作人员的行为数据,还包括心理、安全、工程、人因等各领域专家学者对核电厂研究的科学数据,具有大数据”特征。此外,核能安全的各种研究方法、手段,以及技术都更加多样化,使得人因数据中不仅存在结构化数据,还存在大量的半结构和非结构化的数据,且价值密度低。那么,在“大数据”背景下,如何利用大数据技术解决数据的采集、存储、分析和处理人因数据成为亟待解决的突出问题。

一 核能安全人因大数据的需求

随着对核电厂安全研究的不断深入,累积了大量的人因数据,到目前为止,已经从KB、MB、GB发展到了TB、EB、ZB……数据呈现指数型增长态势,在这海量的数据当中存在很高的参考价值,使人因安全研究结果更趋于精准和可靠。但不同领域的研究人员、不同的研究目标、不同的HRA模型/方法、不同的HRA模型/方法的不同参数需求和数据需求类型等核能安全对人因数据的多样化需求。

(一)不同领域研究人员的人因数据需求。在核电厂,人因研究的主要参与人员有心理学家、HRA专家、工程技术人员等。心理学家主要关心的是人误背景、原因、机制等。HRA专家主要关心的是作业任务、情景状态、关键行为、PSFs、人误概率等。而工程技术人员主要关心的是人误后果、绩效水平等。

(二)不同的研究目标的人因数据需求。例如,在人机系统/人机界面设计方面,需要的人因数据主要有:人体尺寸、生物力学参数、生理负荷参数、心理负荷参数、行为习惯/特性/模式、能力参数、组织管理参数等;在改进系统绩效方面,需要的人因数据主要有:作业模式、关键行为特征、失误模式、失误原因、宏观人因参数等;在安全评价或应用于PSA方面,需要的人因数据主要有:系统特性、任务参数、失

[收稿日期] 2017-12-10

[基金项目] 国家自然科学基金项目“数字化工业系统人因可靠性分析方法研究”资助(编号:71371070);湖南省自然科学基金项目“基于人因可靠性的核电厂数字化主控室人机界面布局优化模型研究”资助(编号:14JJ7046)

[作者简介] 王以群(1959-),女,湖南长沙人,南华大学管理学院教授。

①南华大学管理科学硕士研究生。②湖南工学院人因研究所副教授,博士。

误模式、PSFs、人误概率、人误机制等^[3]。

(三)不同 HRA 模型/方法人因数据需求。概率安全评价中,人因可靠性分析技术研究 HRA 模型/方法有:人误率预计技术(THREP)、人的认知可靠性模型(HCR)、认知可靠性与失误分析(CREAM)^[4]等几十种方法,根据情况的不同需要采用不同的分析模型和方法。在人因研究中,数据需求分为定性数据和定量数据。

(四)不同的 HRA 模型/方法有不同的参数需求。THERP 模型中主要参数有:工作任务、PSFs、人-机界面、工作成员与组织、任务时间窗口、规程、恢复因子、相关性、基本人误概率;HRC 模型中主要参数有:作业任务的任务和行为、PSFs、认知水平、人-机界面、时间窗口、诊断时间、执行时间等;ATHEANA 模型中主要参数有:工作任务和行为、不安全行为、失误迫使情景(EFCs)、EFCs 概率等。

二 人因大数据处理面临的问题

(一)海量、低价值密度人因数据的集成与抽取

核能安全人因数据涉及面较广,来源众多,数据量大且价值密度低,必须从中把人因数据抽取出来,进行集成、分析和解释。例如需要从大量的核电厂监控录像中提取人因失误的片段,分类集成,然后用人因分析技术进行数据分析,并使用规范的格式解释人因数据。而人因数据的海量、低价值密度特征加剧了数据采集的困难,单纯依靠人工采集是不可能完成的任务,且易造成数据采集过程中采集效率低下、人为误差较大等问题,必须借助于大数据技术。可见,海量、低价值密度数据的抽取、集成、分析、解释问题将是大数据人因数据采集、分析与处理面临的第一个挑战^[5]。

(二)非结构化数据的处理与融合

核能安全人因研究的科学研究更多的是数据驱动的、团体协作的和跨学科领域的。这些数据的采集难以在一个统一的框架下进行,容易存在表达存在差异、数据类型和文件格式不同等异构及信息重叠问题。不同途径所获取的人因数据类型是不同的,但是可能是对同一对象的描述或不同角度的解释,而数据可能有不同的物理位置和存储方式,人因研究专家就需要访问很多不同的数据源,这其中存在对人因数据进行非结构化数据的处理和对语义异构信息有效融合的问题。

(三)新型数据的表示方法与关联关系

数据获取与采集新技术手段如眼动仪^[6]带来了新的人因数据类型,这些新型数据无法使用传统

的数据表示方法。如何分析和处理这类实时动态数据的关联关系与因果关系,以支持实时的处理与交互操作,是当前人因大数据处理必须解决的关键问题。

(四)结合人因安全分析技术的数据挖掘算法

对于海量、类型庞杂的超高维人因数据,通过数据降维、大规模并行处理,如何借助大数据技术,结合人因失误辨识技术等入因安全分析技术对人因大数据处理,运用合适的数据挖掘算法建立接口,将是人因数据处理技术面临的挑战。

三 人因大数据采集与处理系统框架

(一)系统框架

应对前面提到的问题,以满足核能安全预防与保障人因数据分析各种需求,针对人因大数据的特性,运用大数据处理技术,建立核能安全人因数据采集、存储、分析与处理系统框架,其结构如图1所示。在该框架中:

1.数据源层,其功能是海量、低价值密度人因数据的集成与抽取,满足不同领域和不同分析目的人因数据需要。通过尽可能广泛的、多样人因数据采集途径,如核电厂运行数据(系统值班日志、设备记录、维修报告等)、人误数据库、WANO、模拟实验、仪器测量数据等进行人因数据采集,将采集到的结构化和非结构化数据,用分布式架构完成对海量人因大数据的存储和数据管理,从而形成人因数据数据库。

2.数据预处理层,其功能是非结构化数据的处理与融合,满足不同的 HRA 模型/方法的人因数据需求。通过对采集到的人因大数据进行真实性、完整性判别,数据清洗过滤,数据结构化处理和索引标记,格式化分类等等,建立多源、异构、非完整、非一致、非准确数据的集成与接口。

3.人因大数据处理中心,利用可视化分析、数据统计、数据挖掘、文本分析、图形分析、空间信息分析、智能语义分析等技术对预处理过的人因数据进行聚类分析与主题提取,数据挖掘与决策支持,满足各领域专家和 HRA 研究模型对人因数据的多样化需求。

4.大数据交互系统,利用新型数据的表示方法与关联关系,为用户提供统一访问接口,满足核能安全人因分析信息数据显示与查询,实时交互集成的需要,为核电安全提供决策支持。

(二)数据源层

人因数据来源是多方面的,而人因事件是整个

平台组织数据的基础,围绕人因事件分析角度、时间维度、数据类型、人因分析模型,结合人因分析流程,在图 1 架构中,把人因数据来源主要分为核电厂

DCS 系统、模拟仿真平台、人误数据库以及人因实验,且分别建立对应数据库。

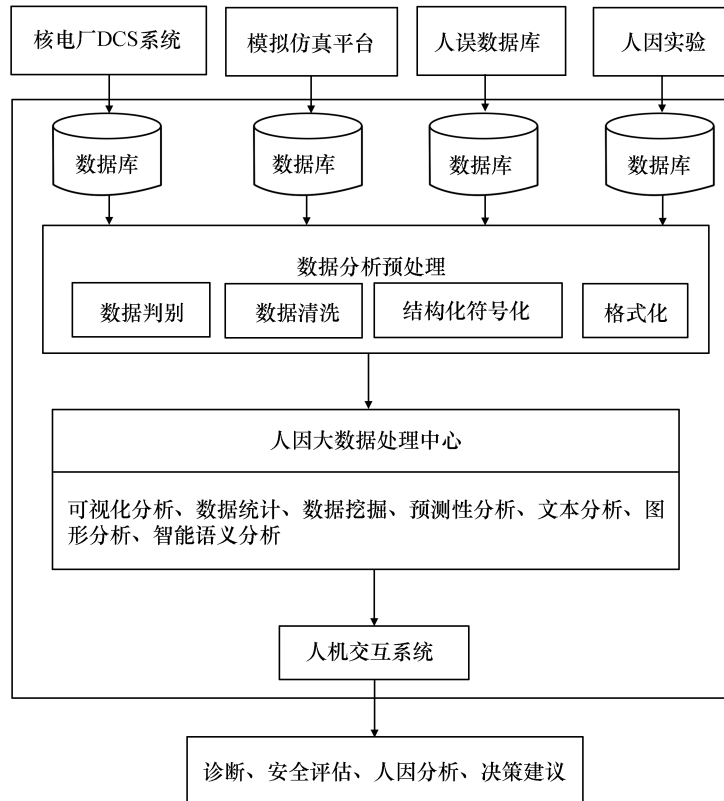


图 1 人因大数据采集与处理系统架构

1.核电厂 DCS 系统数据源,包括核电厂/核设施事件/事故报告/运行经验反馈,系统值班日志、设备记录、维修报告,人误数据库,如 THERP 手册,模拟机实验,实验室实验,访谈,专家判断,文献资料,如 PSA 报告等。通过这些途径得到的数据可信度高,真实可靠,但是存在数据量不足,有效数据极少,分析工具缺乏、分析难度大等采集困难。

2.模拟仿真平台,由于运行数据获取困难,通过模拟仿真核电厂发生事故或故障现场运行,采集人因数据。例如,秦山核电厂操纵员可靠性模拟机实验选择包含技能型、规则型和知识型三种认知类型,对电厂运行安全有重大影响的 23 个异常事件(55 个 HIs),对 38 名操纵员事件响应状况和时间进行录像和记录,取得 764 个数据点,经数据处理和分析后获得适合秦山核电厂系统与人员特性的 HCR/HRA 模型基本参数。

3.人因数据库,目前著名的 WANO 人因数据是核能安全人因数据采集的途径之一。它把核电站发生的事故或故障按照主要失误模式进行归类统计,其中包括了未发现报警或者事故征兆、对事故征兆

或事故判断失误、人员之间交流不足/交流不当、操作失误、组织管理不当等。通过对比各类根本原因及原因因子分布状态,得到根本原因与人误模式之间的关联性。

4.人因实验,如眼动仪、脑电仪、多道生理仪、行为捕捉系统、全过程行为监控系统等人因实验,观察和分析 DCS 操纵员的行为、身体状态(如疲劳度)、心理状态(如心理压力),从操纵员的角度采集人因数据,为核能安全研究提供了新的方向。

核能安全人因数据来源增多和技术仪器的发展进步而产生了巨量信息数据储存问题,使人因数据采集和处理的压力增大,制约着核能安全人因研究的发展。通过数据压缩技术能减少数据计算量,提高存储效率。数据压缩虽然减少了存储空间,但是压缩和解压缩造成了大量 CPU 资源耗费^[7]。因此必须借助合适的计算与存储平台。核能安全人因大数据可以利用分布式文件系统(如 HDFS、HBase 等)来存储,但为满足快速性要求,必须对系统中的大数据进行分类存储。如,对大量的历史和非结构化数据可以采用 HDFS,对核心研究数据使用

DBase。

就目前而言,获取的人因数据通常以纸质文件和一般格式化的数据形式保存。对于纸质文件,将基于模式识别原理,将人因安全研究所需要搜集的纸质文件分类,研究基于表格的图像特征算法快速提取文件模版的图像特征,并在数据库构建对应的数据表格。构建图像采集硬件系统,自动识别输入的纸质文件类型,进行快速特征匹配和图像注册,采用文字识别技术识别特定区域的文字信息,结合人智最终审核将数据自动加载到数据库系统。对于一般格式化的数据,将根据数据类型进行分类并基于数据特征进行处理。这能够大大提升数据录入数据库系统的效率,对于快速增加数据规模、提升数据价值密度、深入获取数据信息有重要帮助。

(三)数据预处理

人因大数据是建立在多数据来源的基础上的,不同数据来源有不同的表达方式,不同科学研究方式、角度也产生了不同的人因数据结构。为解决人因大数据多源、异构、非完整、非一致、非准确数据的集成、分析和解释问题,必须对人因大数据进行预处理^[8]。为此,在数据预处理阶段是对数据进行数据判别、数据清洗、结构化符号化和格式化。首先,从人因分析的角度判别数据是否语义完整、准确和可用。然后过滤掉非完整、非一致、非准确的数据。紧接着,以人因研究对象的基本事件为中心,以研究对象的心理因素、生理因素、组织因素等为层次,在不同层次的不同分析方法和取得的数据资料元数据。最后,对于不同的数据源建立统一的数据索引结构分类,在这个过程中首先要完成人因数据的标准化与规范化工作。

为了适应人因数据的多样化需求,通过 JDBC、Web Services、FTP 等方法实现异构系统与异构数据对接^[9],运用 XML 技术和智能语义分析技术进行数据处理和标记,采用 H-Tree 方法建立索引服务体系,实现批量化的数据统一适配器。通过对异构数据面向 HRA 分析模型和各专家需求,建立标准事件模型以及标准数据元映射服务,使数据转化为标准化结构化的数据集合。通过调用建立的索引服务,实现核能安全人因大数据管理的标准化与规范化,完成人因数据的整合与归集。

(四)人因大数据处理中心

人因大数据处理中心的主要任务是根据人因分析需要。对预处理人因数据进行聚类分析与主题提取,数据挖掘与决策支持。按照人因数据分析逻辑,首先完成基于对象基本事件的人因分析逻辑配置。

依据心理因素、生理因素、组织因素等层面分别进行面向人因事件的分析逻辑配置,实现符合人因分析思路的参数配置。其次,按所需要的不同分析断面,分别对涉及的关键人因事件进行数据挖掘和分析。使用 MapReduce 分布式计算技术,根据文件类型选用不同的人因数据处理工具和模型方法,运用数据挖掘技术分别进行并行计算,完成符合人因分析需要的人因因子提取和配置。最大限度发挥人、机各自优势的人机交互与最优化协作,完成对基于人因分析模型的高效分析。

常见的数据挖掘算法有神经网络(包括 BP 网络和 RBF 网络)、决策树、粗糙集、贝叶斯、遗传算法、Apriori 算法、PageRank 算法等^[10]。其中 Apriori 算法是经典的布尔关联规则算法。在分类上,其关联规则属于单维、单层、布尔关联规则。该算法是从数据中发现频繁模式的有效算法^[11]。人因数据分析主要是发现布尔型的关联规则,因而 Apriori 算法是适合于人因数据的数据挖掘算法。

Apriori 算法也被称为购物篮分析(Market Basket analysis),于 1994 年由 Rakesh Agrawal 和 Ramakrishnan Srikant 提出。其步骤主要包括候选集生成和迭代剪枝这两个阶段。该算法的主要步骤为:

(1)生成候选集:找出所有出现的频繁性至少和预定义的最小支持度一样频集。

(2)剪枝:由频集产生满足最小支持度和最小可信度的强关联规则。

(3)递归迭代:采用中规则,使用第(1)步找到的频集产生只包含集合的项的所有规则,其中,每一条规则的右部只有一项。最后筛选出大于用户给定的最小可信度的规则。

Apriori 算法的缺点在于:可能会产生大量的候选集和需要重复扫描数据库,大大增加了执行时间,降低效率。为了解决这个问题,很多人做了尝试,如 Savasere 等设计了一个基于划分的算法,这个算法解决了重复扫描数据库的问题,但是其通信过程增加了算法执行时间。J.Han 等提出了 FP-树频集算法,该算法不产生候选挖掘频繁项集,不仅能很好适应不同长度的规则,还显著提高了效率。

对于人因大数据处理中心,由于数据面涉及广,数据量大,类型繁多,因此,数据的价值密度不高,采取数据挖掘技术是发现大量数据有效价值的技术^[12]。对数据管理系统数据库基于 Apriori 算法进行关联规则自动提取,对于时间序列数据进行时序规则提取,建立关联规则库。根据建立的人因失误模型和机理提取关联规则,设计关联规则集优化算

法,消除关联规则集中的不一致。最后可以利用 MapReduce 的快速处理性能进行自动化批量处理。

(五)人机交互系统

为增加系统易用性和满足应对不同领域专家需求,必须实现数据可视化,建立统一的、标准化的数据访问接口,使显示界面符合人的认知特征^[4],达到最佳显示的目的。PSA 研究领域经过几十年的发展,积累了大量各具特色的人因数据表征,这将为解决人因大数据采集与处理系统的可视化问题提供有力的支持。基于人因大数据中心形成的事故案例资源子集,建立灵活的人因分析断面,支撑核能安全研究分析断面数据的显示。分析结果数据经封装后使用统一的标准和模式,用可视化技术处理后输送到客户端,进行人因分析报告展示。

四 结束语

人因数据是核能安全研究的重要数据来源和组成部分,解决人因大数据问题是核能安全领域人因研究的重要环节。作者在此仅构建了核能安全人因数据采集、分析与处理系统框架,试图运用大数据技术来解决人因大数据所带来的问题,更加快速、及时、精确地应用先进的 HRA 方法和 PSA 方法对核电厂的安全进行评估提供帮助。至于框架中所涉及的人因大数据处理技术是今后的研究重点,作者拟从数据挖掘技术入手,对所收集和储存的人因大数据进行处理,从而完善该系统框架的可用性,其详细的研究工作有待日后与学者交流。

[参考文献]

[1] 方巍,郑玉,徐江.大数据:概念、技术及应用研究综

述[J].南京信息工程大学学报(自然科学版),2014,6(5):1-2.

- [2] ROGERS M L, SOCKOLOV P S, BOWLES K H, HAND K E, GEORGE J. Use of a human factors approach to uncover informatics needs of nurses in documentation of care [J]. Med Inform, 2013, 82(11): 1068-1074.
- [3] 戴立操,张力,黄曙东.复杂工业系统中人因失误根本原因分析[J].中国安全科学学报,2003,11(11): 13-16.
- [4] 蒋建军,戴立操,张力等.基于隐马尔可夫的核电厂半数字化人—机界面事故诊断过程人因可靠性模型[J].核动力工程,2012,33(5): 80-82.
- [5] 任磊,杜一,马帅,等.大数据可视分析综述[J].软件学报,2014(9): 1909-1936.
- [6] KIM S K, SUH S M, JANG G S, HONG S K, PARK J C. Empirical research on an ecological interface design for improving situation awareness of operators in an advanced control room[J]. Nucl Eng Des, 2012(235): 226-237.
- [7] 李佳伟,郝悍勇,李宁辉.电网企业大数据技术应用研究[J].电力信息与通信技术,2014(12): 20-25.
- [8] 白如江,冷伏海.“大数据”时代科学数据整合研究[J].情报理论与实践,2014,37(1): 94-99.
- [9] 李维,陈祁,张晨,等.基于大数据技术的临床数据中心与智能分析应用平台构建[J].医学信息学杂志,2014,35(6): 13-17.
- [10] 李平荣.大数据时代的数据挖掘技术与应用[J].重庆三峡学院学报,2014(3): 45-47.
- [11] 樊嘉麒.基于大数据的数据挖掘引擎[D].北京:北京邮电大学,2015: 24-29.
- [12] 张力,陈文,等.基于 DPSIR 和 BP 神经网络的安全绩效评估模型[J].中国安全科学学报,2014,24(12): 77-82.

Framework of Nuclear Safety Human Factors Data Acquisition and Processing System in the Context of Big Data

WANG Yi-qun, WU Bo, FANG Xiao-yong

(University of South China, Hengyang 421001, China)

Abstract: An analysis of human factor is becoming more and more important in nuclear safety management. However, there are different human factors data requirements for different human factors analysis fields, targets and methods. At the same time, in the context of big data, human factors data analysis faces the integration and extraction of massive and low-value human data, unstructured data handling and integration issues. In order to solve these problems and challenges, this research sets up the system framework of human factors data on the basis of features of human factors data collection and processing, which consists of four aspects including data source layer, data preprocessing, human factors big data processing center, and human-computer interaction system.

Key words: nuclear safety; human factors; data acquisition and processing; Big Data technologies