WANO 人误研究平台①

郑 龙[1], 张 力[1,2]

[1](南华大学 人因研究所, 衡阳 421001)
[2](湖南工学院, 衡阳 421001)

摘 要: 针对传统 WANO 人误研究工作效率低下, 流程混乱的特征, 研究从 WANO 人误研究工作实际需求出发, 采用面向对象技术设计, 实现了一个基于信息系统的 WANO 人误研究平台. 从平台构建的角度阐述了 WANO 人误研究平台的功能结构、平台框架、WANO 人误数据库构建等技术要点, 并介绍了平台在 WANO 人误数据 ETL 方面的应用效果.

关键词: WANO 人误研究; 面向对象; 研究平台; 人误数据库

WANO Human Error Research Platform

ZHENG Long[1] , ZHANG Li[1,2]

[1](Human Factor Institute, University of South China, Hengyang 421001, China)
[2](Hunan Institute of Technology, Hengyang 421001, China)

Abstract: In view of the inefficient and chaos of traditional WANO human error research work, the study based on actual demand of the WANO human error research work and object-oriented technology. It designed and implemented a WANO human error research platform. The article from the perspective of the WANO human error research platform construction elaborated platform realize related technical key points, such as platform framework, WANO human error database construction etc, and introduced the application effect of data ETL on WANO human error research platform .

Key words: WANO human error research; object-oriented; research platform; human error database

WANO 人误研究[1]是一项基于世界核电运营者联合会发布的事件报告[2]进行数据的萃取, 转置和加载(ETL)操作, 并在其基础上进行数据挖掘相关的统计分析研究[3], 以跟踪和发现核电领域人因失误的规律和发展动态的研究工作[4]. 研究的主要数据来源是 WANO 定期发布的 Html 网页格式的事件报告. 然而, 由于 WANO 每年发布的事件报告多达上千份, 而且都是非结构化的, 把其转换为满足研究需要的结构化数据, 传统的处理方式需要大量的时间和人力去进行数据的萃取和整理, 并且数据只能保存和显示为单一格式的 Excel 文档, 缺乏弹性和可扩展性, 在其基础上进行统计分析工作效率低而且可视化程度低. 并且由于 WANO 人误研究是一项长期的研究工作, 长期以来, 积累了大量的 WANO 事件报告数据, 这些数据具有获取成本高、原始而唯一、不可再生和可重复利用的特点,需妥善保管和整理, 以备未来研究中查阅使用. 对这些资料应有效地管理、研究和利用, 避免重复工作, 提高工作效率, 节省成本. 然而, 传统的方式无论是在 WANO 人误数据收集、人误数据维护、人误数据准备方面, 还是人误数据分析分析方面都存在效率低下, 资料保存分散, 综合研究不够, 信息化程度低, 不利于资料的交流和共享, 不能满足当前人误研究工作的需求, 更不能充分挖掘 WANO 事件报告数据的潜在价值[5]. 因此构建一个既能充分满足当下 WANO 人误需求, 又有助于人误研究工作的长期合理发展的 WANO 研究平台显得尤为重要. 研究平台作为一种基于信息技术实现的面向研究应用的信息平台, 其能够为研究工作提供各种基础支持(数据、组件、模板等), 并且能集成研究工作的各种应用需求, 使研究成为一个完整的体系, 成为提升研究工作效率和规范性的一种有效

① 收稿时间:2014-08-23:收到修改稿时间:2014-10-20

手段, 获得了广泛认可和应用. 譬如: 夏菁, 白志强等设计实现的地质资料信息化综合管理平台, 依据专家经验设计了平台总体架构和主要功能模块, 搭建了地质资料数据库, 实现了地质数据管理与维护、地质图件管理与维护、综合信息查询、用户互动交流、数值分析、数据图表和系统安全管理等功能[6];张晓梅, 陈旭等设计和开发的指纹识别算法研究辅助平台. 该平台根据工作流思想将算法体系抽象成算法链路模型, 使指纹识别算法同平台之间形成松耦合, 便于新算法集成到平台, 以支持指纹识别研究工作[7].

本研究以 WANO 人误数据挖掘、维护、准备和分析作为平台功能和流程设计的主线, 完成了 WANO 人误研究平台功能模块的设计和架构设计, 基于 WANO 人误研究对数据的需求, 设计实现了平台数据库(WANO 人误数据库), 并介绍了平台关键的 Html 文档解析功能的实现方式和人误数据 ETL 应用效果.

1 平台功能模块及总体框架设计

WANO 人误研究平台采用面向对象和模块化的设计方法, 平台提供基础的数据支持, 功能组件和应用模板支持, 用户可以根据自身的应用需求, 灵活调整功能组件,开发和调用不同功能模板,从而在降低程序复杂度, 便于平台设计和实现的同时提高平台的适应性和可操作性[8]. 平台构建的总体目标是实现一个能够支持 WANO 人误研究的集数据采集, 数据维护, 数据准备和数据统计分析功能于一体的研究平台. 通过对 WANO 人误研究的应用需求的获取和分析, 从整体上形成了几条平台构建思路:

(1) WANO 人误研究平台限于研究所内部使用, 用户群体数量有限、身份明确、素质较高, 易于组织培训和管理. 因此平台可以设计得更具弹性和复杂性, 从而在最大程度上满足用户多维需求, 提高平台的适应性和可用性.

(2) 平台采用开放, 易扩展的功能结构, 基于这个平台, 用户可定制自己需要的数据视图和数据应用功能, 主要体现为: ①数据应用模块的功能主要通过组件和模板的调用来实现, 在统一应用平台之上, 用户可以根据自己需求定制,开发应用组件和模板, 调用组件和模板实现相应的功能. ②数据源管理是另一个平台开放性, 可定制性的体现. 通过数据源管理模块, 用户可以根据自身需要, 基于 WANO 编码体系定制自己的数据视图.

(3) WANO 事件报告中的内容是基于 WANO 编码体系的, 为了无损的以结构化的方式保存 WANO 事件报告数据信息. 平台后台数据库需以 WANO 编码体系为基准进行设计, 以支持人误分析和规则挖掘.

1.1 功能模块设计

WANO 人误研究平台的功能设计和模块划分按照功能性质的不同, 分为 WANO 人误研究业务功能和公共辅助功能两大功能区, 其中公共辅助功能区由一些起基础支持功能的模块组成, 业务功能区则以 WANO 人误研究工作中人误数据的收集、维护、准备和应用为主线划分为四大功能模块. 如图 1 所示.

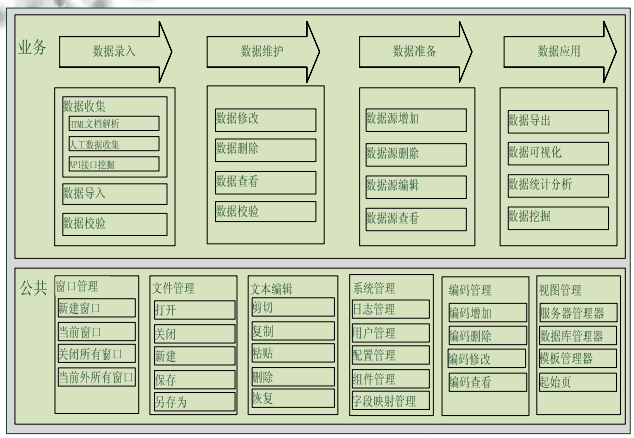

图 1 平台业务功能

公共辅助功能区由六个子功能模块组成, 着重提高平台的可操作性和用户体验, 为 WANO 人误研究业务功能的实现提供基础支持, 其中, ①窗口是用户与平台交互的基础, 窗口管理模块提供对平台应用窗口的统一管理. ②系统管理模块提供对平台配置的全面管理, 主要由五部分组成: 日志管理、用户管理、配置管理、组件管理和字段映射管理. ③WANO 编码体系是 WANO 人误数据库的基础, 编码管理模块提供对 WANO 编码体系的全面维护. ④视图管理模块为用用户提供对平台功能多维视图, 通过服务器管理器平台可以实现对数据库服务器的管理; 数据库管理器提供对服务器下所有数据库及其数据库子对象的管理; 模板管理器是提供平台功能扩展的关键, 通过其可以对基于平台接口开发的应用模板进行管理, 从而满足平台功能扩展和用户定制化需求.

业务功能区提供对 WANO 人误研究的业务功能

的主要支持, 其由四个功能模块组成, 其中, ①数据录入功能提供对 WANO 人误数据收集的全面支持, 主要表现为三个方面: 数据收集针对的是非格式化 WANO 人误数据的收集, 其中 HTML 文档解析和人工数据收集提供了两种从 WANO 事件报告中提取有用数据的方法, API 接口挖掘功能基于 WANO 提供的应用程序接口, 为从 WANO 后台数据库挖掘人误数据提供支持. 数据导入则是基于研究所长期 WANO 人误研究过程中收集起来的 Excel 文档数据, 提供格式化人误数据的批量导入; 数据校验功能是把收集得到的数据同 WANO 编码体系进行匹配, 标识出匹配失败的数据项, 为数据处理提供支持, 只有校验通过的数据才能存入 WANO 人误数据库. ②数据维护功能模块为 WANO 人误数据库中的数据提供全面的管理功能, 并为 WANO 数据提供安全支持, 只有通过数据校验的数据才能反应到研究应用结果中. ③数据准备模块用于对数据指标进行操作和管理, 其本质是一种元数据管理, 通过这些元数据, 平台应用可以从数据库得到定制化的数据视图, 在其基础上对平台数据进行操作, 平台中表述为数据源管理主要是因为其是统计分析、数据挖掘、数据可视化等应用功能的数据来源. ④数据应用模块可以在数据源和功能组件的支持下对 WANO 数据进行各种操作, 包括数据统计分析、数据格式化导出、数据可视化等功能.

1.2 架构选择

结合平台构建需求和三层架构的优缺点[9], 本研究选用数据层驱动模式即数据层构建为中心的模式, 并 B/ S 与 C/ S 结构模式各有自己的优点与不足之处, 根据取长补短、优势互补的原则, 集成两种结构模式来构建 WANO 人误研究平台. 如图 2 所示.

图 2　平台架构

研究根据各个功能定位的不同, 把平台总体架构分四个层级, 其中, ①数据持久层是独立于业务功能之外的数据支持层, 其为平台业务功能实现提供数据支持. 在本平台中, 主要有六种方式向平台提供数据存取支持, 分别是高度结构化的 WANO 人误数据库、格式化文档(Excel、Txt 等文档)、非格式化文档(Html 文档)以及 Xml 格式的字段映射文档、平台配置文件等. 其中平台配置文件、应用配置文件、字段映射文件主要存取平台的配置信息; 数据库配置文件则是平台数据库服务器和数据库连接信息的存取载体. ②数据访问层作为三层架构的底层, 提供了对数据持久层的访问功能, 通过其可以为业务功能层的业务逻辑实现和外部接口提供数据支持. ③业务功能层在三层架构中起到承上启下的作用, 通过业务功能层用户可以基于数据访问层定制自身所需的功能和服务. 在本平台中业务功能层主要由两部分组成, 业务逻辑类库提供了许多平台内置的功能和服务, 如一些简单的统计分析功能、数据挖掘功能、数据可视化功能等; 外部接口类库主要是提供了一些访问其它一些应用程序的接口, 如 SPSS、SAS、EXCEL 等, 借用其强大的功能实现平台功能的扩展. ④用户访问层, 顾名思义就是以可视化的方式向用户提供平台功能访问和调用的 UI 接口, 向用户提供体验良好的界面和服务视图. 基于 WANO 研究的现实需要及 B/S 和 C/S 各自的特点, 平台在相同的数据访问层和业务功能层的基础之上, 为用户提供了 B/S 和 C/S 两种用户访问方式. 平台中 B/S 模式主要提供数据浏览功能, C/S 模式主要提供功能和服务的调用功能. 如图 2 所示, WEBForm 和 WINForm 两种表现控制类在服务器端进行服务调用以及对页面初始化, 通过数据推送服务层与客户端进行通信, 最后把结果呈现给用户, 实现平台与用户的交互.

2　平台实现

WANO 人误研究平台是基于.NET 4.0 和 SQL2008 实现的, 主要涉及到 WANO 人误数据库设计与实现和 HTML 文档解析功能实现两个方面.

2.1 WANO 人误数据库构建

数据库在信息平台中占有非常重要的地位, 数据库结构的好坏将直接对应用平台的效率及实现的效果产生影响, 合理的数据库结构设计可以提高数据存储的效率, 保证数据的完整性、一致性和安全性并降低

平台应用开发的难度.

本研究基于 WANO 编码体系和对实际应用的需要的分析构建起概念模型, 采用数据包图以数据打包的方式反映主题数据的多维性. 如表 1, 以 WANO 人误事件报告主题数据为例, WANO 人误数据包图以二维表的形式表示了 WANO 人误数据主题相关的多维数据视图, 如事件发生日期、事件所属类型、发生事件的反应堆类型、事件相关的人员、这些视图在提高数据规范, 降低数据库冗余的同时为用户提供了多个对 WANO 事件报告进行研究的视角.

表 1 WANO 人误数据包图

事件日期维	事件类型维	反应堆类型维	界面类型维	国度维	报告状态维	制造商维	相关人员维	故障设备维	事件因素维	直接原因维	非直接原因维	事件后果维
年份	编码	编码	编码	编码	编码	编码	一级编码	一级编码	一级编码	一级编码	一级编码	一级编码
月份	名称	名称	名称	名称	名称	名称	二级编码	二级编码	二级编码	二级编码	二级编码	二级编码
日期	状态	状态	状态	状态	状态	状态	三级编码	三级编码	三级编码	三级编码	三级编码	三级编码
	备注	备注	备注	备注	备注	备注	状态	状态	状态	状态	状态	状态
							备注	备注	备注	备注	备注	备注

逻辑模型主要是体现 WANO 人误研究业务方面的需求, 对平台后续的物理模型设计和实施起指导作用. 其主要工作有: ①分析主题域, 即依据研究需要确定对不同的数据维进行组合, 譬如: 分析职能与人误事件发生的关系时, 就应从相关人员、事件因素、事件后果等数据维度进行查询分析. ②确定数据粒度. 所谓粒度就是划分数据库中数据单元的详细程度和级别的一种量级, 数据详细粒度小, 数据简单粒度大. 平台应依据研究对数据的需求, 确定数据粒度. 譬如: 对于综合性很高的数据, 就应选用较大的数据粒度, 以减少研究时的数据量. ③数据分割, 即按照一定的原则(如业务需求、时间、类型等)对粒度太大的数据进行分割, 置于不同的物理单元中, 使这些数据既可单独使用, 也可组合使用. 从而提高存储效率和可用性.

WANO 人误数据库物理模型设计是基于 Sysbase 公司的 PowerDesigner 辅助设计软件实现的, 其任务是将逻辑模型转变为实际的数据库(SQL2008)存储, 其主要内容包括: 确定数据库对象的命名规范、建立数据库的物理模型、确定数据库的索引策略等. 最后通过运行由 PowerDesigner 导出的数据库脚本实现 WANO 人误数据库.

2.2 HTML 文档解析功能

WANO 人误统计分析平台的原始数据来源 WANO 发布在其官方网站的 HTML 网页格式的事件报告, 如何从这种格式的文件中提取有用数据是本研究的重点之一. 但是研究发现, 运用.NET Framework 类来解析 HTML 文件,读取数据并不容易[10]. 虽然可以用.NET Framework 中的许多类(如 StreamReader)来逐行解析文件, 但 XmlReader 提供的 API 并不是“取出即可用(out of the box)”的, 这是因为 HTML 的格式是不规范的[11]. 通过探索, 本研究最终选用了由 Microsoft 的 XML 大师 Chris Lovett 发布的一个 SGML 解析器, 叫做 SgmlReader, 通过它可以解析 HTML 文件, 甚至将它们转换成一个格式规范的结构. SgmlReader 本身派生于.NET Framework 中的 XmlReader 类, 也就是说, 可以像运用诸如 XmlTextReader 这样的类来解析 XML 文件那样来解析 HTML 文件并生成格式规范的 HTML, 从而可以运用 XPath 语句来读取数据[12]. 譬如: 可以通过如下步骤通过一个 URL(统一资源定位符)实现对一个远程 HTML 文件的解析:

(1) 通过实例化 HttpWebRequest 和 HttpWebResponse 对象来访问一个远程的 HTML 文件:

```
HttpWebRequest req =(HttpWebRequest)WebRequest.Create(uri);
HttpWebResponse res =(HttpWebResponse)req.GetResponse();
StreamReader sReader = new StreamReader(res.GetResponseStream());
```

(2) 创建一个 SgmlReader 类, 并把它的 DocType 属性设置为“HTML”:

```
SgmlReader reader = new SgmlReader();
reader.DocType = "HTML";
```

(3) 把前面加载有 HTML 文件响应流的 sReader 对象赋值给 SgmlReader 类的 InputStream 属性, 从而把 HTML 文件流加载到 SgmlReader 对象:

```
reader.InputStream = new StringReader(sReader.ReadToEnd());
```

(4) 然后可以通过调用 SgmlReader 的 Read()方法来解析 HTML 文件并创建格式规范的 HTML 文档:

```
sw = new StringWriter();
writer = new XmlTextWriter(sw);
writer.Formatting = Formatting.Indented;
while (reader.Read()) {
if (reader.NodeType != XmlNodeType.Whitespace)
```

```
{writer.WriteNode(reader, true);
    }}
```

(5) 最后就可以用 XPath 语句来读取 HTML 文档中不同的节点, 从而实现 HTML 文档的数据读取, 下面的代码示例了如何将 SgmlReader 生成的输出结果加载到一个 PathNavigator, 然后使用 XPath 语句来查询 HTML 文件结构:

```
StringBuilder sb = new StringBuilder();
XPathDocument doc = new XPathDocument(new StringReader(sw.ToString()));
XPathNavigator nav = doc.CreateNavigator();
XPathNodeIterator nodes = nav.Select(xpath);
while (nodes.MoveNext()) {
sb.Append(nodes.Current.Value);}
return sb.ToString();
```

3 应用效果

WANO 人误研究平台已经实现了数据录入、数据维护、数据准备和公共功能辅助区的大部分功能, 目前已经可以为 WANO 人误研究工作提供基础的数据管理功能, 应用测试表明平台的应用很大程度上提高了研究工作的效率. 以 WANO 人误数据管理方面的功能为例: 数据录入整合了以前存在的多种数据来源, 借助 Html 页面解析功能平台平均每小时能处理 50 份事件报告, 总体的准确率达到了 96%, 原有的基于人工 WANO 编码体系对照数据收集方式, 每小时只能处理 6 份事件报告, 准确率只有 90%, 处理效率和准确率都得到了较大的提升; 数据维护方面, 平台提供了全套的数据维护功能和安全控制功能, 提高了数据的准确性和安全性. 传统的数据维护完全依靠个人, 缺乏规则的限制和有效的数据安全保护, 数据准确性的安全性毫无保障; 数据准备方面, 平台为用户提供十分便捷的数据视图定制功能, 用户一般可以在十分钟内得到自己所需的数据, 而传统方法在源数据完整的情况下, 也往往需要一两天的时间, 经过复杂的操作, 花费大量的精力才能筛选出自己所需的数据. 相信, 随着数据应用功能模块的实现, 平台将为 WANO 人误研究工作带来更多便利.

4 结语

本文从 WANO 人误研究工作的实际需求出发, 对 WANO 人误研究平台进行了功能定义和模块设计. 并对系统的框架, 进行了探索研究, 提出了一个适于本系统构建需求的框架结构. 同时也对平台实现十分关键的 WANO 人误数据库设计与实现, 组件管理和 HTML 网页解析技术进行了介绍. 并基于已经实现的平台功能, 从数据录入、数据维护、数据准备三个方面与原有的研究方法进行了效率对比, 表明 WANO 人误研究平台能在很大程度上提高 WANO 人误研究效率. 目前 WANO 人误研究平台的开发工作还未全部完成, 但基于前期的研究成果, 开发工作正在稳步有序的推进. 相信该平台的实现和投入使用定能为 WANO 人误统计分析工作的规范化和效率提升做出重要贡献.

参考文献

1 高佳,沈祖培,黄祥瑞.人的可靠性分析:历史、需求和进展.中国安全工程学报,2003,(12):44–47.

2 World Association Nuclear Operators. Event Reports. http://www.wano.org.uk.

3 Wang Z. Strategic value and enlightenment to promote the development of U.S. big data technology. China Development Observation, 2012, (6): 44–45.

4 赵明,张力.数据挖掘在人因事件分析中的应用.国际安全科学与技术学术研讨会论文集.2008.93–97.

5 张力,赵明.WANO 人因事件统计及分析.核动力工程,2005,6(3):291–296.

6 夏菁,白志强,王宝鹏,常洁琼,吴一超.地质资料信息化综合管理平台的设计及实现.北京大学学报,2014,3(2):295–300.

7 张晓梅,陈旭,任春晓,尹义,詹小四.指纹识别算法研究平台的设计与开发.计算机工程与设计,2008,7(14):3083–3089.

8 Zhao F, Liu JJ. Application estimation of Commercial information of cloud computing in system. Communications Technology, 2012, 45 (4): 7–9, 12.

9 高文宇,张力.人因可靠性数据库基础架构研究.中国安全工程学报,2010,12(12):63–67.

10 Matthew A. Russell Mining the Social Web. O'Reilly Media, Inc., 2010.

11 蒋琴琴,宫哲,辛阳.基于 HTML Parser 的 BBS 信息抽取系统的设计与实现.自动化技术与应用,2012,(1):32–37.

12 冯进,丁博,史殿习.XML 解析技术研究.计算机科学与工程,2009,(2):120–124.