文章编号:1671-4598(2018)04-0112-04 DOI:10.16526/j.cnki.11—4762/tp.2018.04.030 中图分类号:TP3 文献标识码:A

基于日志挖掘的装备健康管理系统设计与实现

王国林,介阳阳,叶君好,叶 彬
(中国人民解放军63726部队,银川 750001)

摘要:随着装备的复杂度和信息化程度不断提高,在装备使用过程中产生的日志信息量正以指数级增长;针对目前日志信息处理应用中存在的问题,设计并实现了一个装备健康管理系统,通过对某型装备大量日志文件的分析,制定了设备通用的日志文件预处理规则和重要信息提取适配器,并结合数据挖掘算法构建装备健康管理系统,结合云重心评估法实现了对设备健康状态的评估;该系统从大量日志文件中提取出有价值的信息,帮助设备人员快速定位装备异常,提高解决问题的效率,通过设置系统的关键指标和监测点阀值,实现装备的健康预警。

关键词:日志分析;数据挖掘;健康预警;健康评估;云重心评估法

Designing and Realizing of Equipment Health Management System Based on Log Mining

Wang Guolin, Jie Yangyang, Ye Junhao, Ye Bin
(PLA 63726 Unit, Yinchuan 750001, China)

Abstract: Log information produced from the use of equipment has seen exponential growth with the increasing of the complexity and informatization of equipments. Universal log file reprocessing rules and critical information extracting adaptor are developed through analysis for massive log files of a certain type of equipment, while equipment health managing system is constructed based on data mining algorithm and evaluation for equipment health is realized based on cloud barycenter assessment method. The system can extract valuable information from massive log files and accelerate the location of equipment anomaly, thus can improve problem solving efficiency. It can also realize equipment health warning by configuring critical parameters and monitor threshold.

Keywords: log analysis; data mining; health warning; health evaluation; cloud barycenter assessment

0 引言

日志是指系统对某些对象的某些操作和其操作结果按时间排列的有序集合[1],包含了一个时间戳和一条消息或者系统所特有的其他信息的半结构化数据。每个日志文件[2]由很多的事件记录组成,每条日志记录存储着一次单独的系统事件,能够实时反应系统某一组成部分变化时的信息。日志文件中记录的信息可用于监控系统状态、审计用户操作行为和定位装备异常部位,为解决系统问题提供证据。随着装备的复杂度逐渐增加、信息化程度不断提高、要监控的状态点和参数也越来越多,在装备使用过程中产生的日志信息量正以指数级增长。目前日志信息处理应用中存在以下四个方面的问题:

1)装备在使用过程中产生的大量日志信息缺乏合理的管理和利用。装备在操作使用过程中,日志文件详细地记录了装备软硬件的状态变更信息,这些信息包含着大量的装备状态信息、操作信息、异常信息、参数信息和一些“诡异的”、“在特定环境下”产生的信息。这些日志文件种类多、数据量大、缺乏规范性、可读性比较差、且不同的日志记录之间存在重要的联系。如何对这些日志信息分类存储、管理和利用,对掌握装备的状态起着重要的作用。

2)装备出现问题后的定位模式和依靠纯人工监控装备状态的做法逐渐不能满足使用要求。装备的高复杂度给操作人员使用维护带来了巨大的工作量,需要监控和维护的参数成量级增加,仅仅依靠人员的界面监控已经不能满足此类装备的使用要求。在装备的使用过程中,因为装备的复杂而存在定位问题繁琐、浪费时间等问题。当前,分机人员开展装备的工作模式通常都是在“出现问题”后,逐步排查问题。一旦装备出现故障,分机人员常常需要分析大量冗长的日志来查找装备问题,面对这些大量的日志文件,仅靠手工进行日志分析,效率非常低下,也不便于装备问题的定位。

3)事后排查故障的处理模式逐步不能满足日常用户的需求,需向装备健康预警转变。装备专用软件的日志文件中记录着装备的每一个细节过程,包括装备状态信息、操作信息、异常信息和参数信息等,同时专用软件所在的工控机操作系统也记录着专用软件运行平台的日志信息。通过对这些日志的分析,依据事先设定的关键字异常阀值、构建的系统指标库,当分析结果超出预设阀值或指标范围时,系统会自动告警分机人员,达到健康预警的目的。

4)装备健康状况评估结果对大系统质量的贡献日趋明显。目前,大系统评估依靠人为因素的较多,缺乏靠装备自身产生的信息来评估装备状态的手段。该系统通过对日志文件的分析挖掘,一方面将分析处理结果进行存储作为装备的生长履历,另一方面将其结果结合云重心评估法[3],对装备健康状况进行评估,提供详尽的装备健康状况报告,给系统评估提供参考依据。

为有效解决上述日志处理方面的问题,本文设计和开发了基于日志挖掘的装备健康管理系统。该系统有助于设备操作人

收稿日期:2017-10-30; **修回日期**:2017-12-17。

作者简介:王国林(1976-),男,陕西勉县人,硕士,工程师,主要从事计算机应用技术与装备健康管理方向的研究。

员和总体人员及时分析装备日志，定位装备故障，找出装备的运行规律，实现装备的预防性维修和动态质量控制，达到装备健康预警的目的，对掌握装备的瞬时状态有着重要的意义。

1　系统设计

装备日志文件作为系统的输入，分布于装备不同分系统所在的工控机上，在进行数据预处理前，需建立系统内部各分系统之间的网络通信机制，实现日志文件的网络下载或实时读取，为开展后续的日志统计分析工作做好准备。其主要功能包括：

1）能够自由配置待分析日志文件所在的机器名路径、文件名路径。根据不同的日志类型，构建适用于日志解析的正则表达式。

2）运用正则表达式解析非结构化日志文件，对解析结果进行清洗、规范化处理和存入数据库。

3）应用数据挖掘的分类、关联分析等方法，归纳用户操作与异常之间的关系，找出故障参数或状态与分系统等之间的关系。

4）通过对入库日志的统计分析，找出异常高发的装备软件模块或硬件部件，同时，能够准确定位异常发生的时间和异常所记录的日志文件。

5）根据设定预警的异常阀值，实现对装备健康状况的监控。

6）根据装备指标要求构建指标库和评价指标体系，运用云重心评估法，为装备提供健康状况评估报告。

7）根据数据库中的数据和配置信息生成相应的报表，实现装备履历日志的自动电子化、实现质量控制文档过程的自动电子化，提供直观的结果显示方式。

8）根据系统配置定时生成各种报表，发送给指定的用户或存在于本地文件系统中。提供强大的报表查询功能，用户可以通过时间段、异常代码等多种条件进行查询，并生成直观的报表。

系统的非功能性需求要求系统能够支持不同种类装备产生的日志文件，适应每天产生的大量日志，保证在通过查询条件定位异常信息时，做到高效迅速。系统的总体流程如图 1 所示。

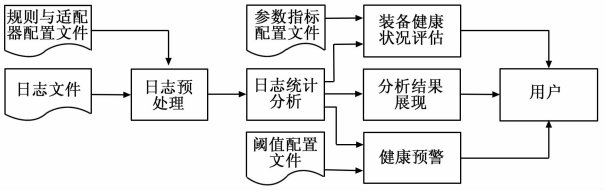

图 1　总体流程图

2　系统实现

2.1　系统功能实现

系统的工作流程主要分为：日志预处理、日志分析、统计查询、装备健康状况评估和健康预警。根据系统的总体流程，下面分别给出不同模块的设计与实现：

1）日志预处理：接收装备的日志文件，针对不同的日志文件，选用不同的正则表达式适配器，读取日志文件，按照配置文件和适配器对原始日志进行解析和过滤，通过分类算法进行统计前的清洗处理，提炼出有用信息，将无关的噪声信息去掉[4]。其流程如图 2 所示。

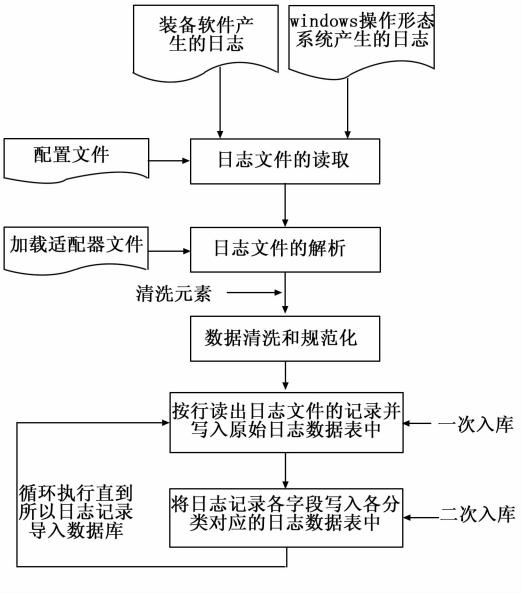

图 2　日志预处理流程图

2）日志统计分析：负责对清洗后的日志文件进行再次分类过滤，合并聚类统计[5]，将生成的分析统计结果记录到数据库中。

3）分析结果展现：按照查询条件对日志的统计结果进行查询，将返回的统计数据按照报表文件的要求生成相应的报表，以人性化的 UI 展现给用户，同时，积累的日志分析数据直观的反应了装备生长健康状况。

4）装备健康状况评估：为了适应视情维修[6]，改变早期的“事后维修”和“定期维修”体制。构建装备健康评估指标体系和评估方法模型，把日志中记录的装备工作状态、产生的各种反应装备状况的数据和各种统计分析结果作为输入带入评估模型，一方面对装备的健康状况给出评估，另一方面，诊断和预测装备未来可能发生的故障，判断其健康状态的好坏，得出其发展规律，从而制定合理的装备维修计划，提高装备运行的可靠性、安全性和有效性。其流程如图 3 所示。

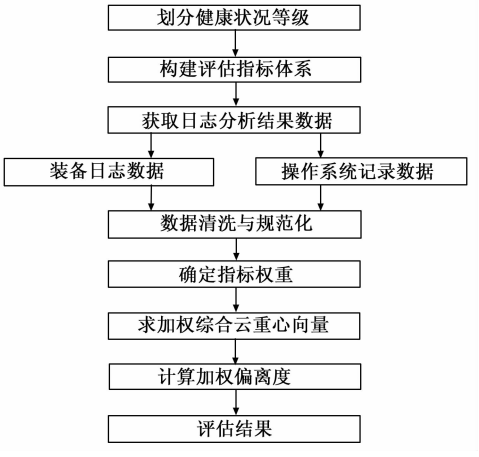

图 3　装备健康状况评估流程图

5）健康预警：依据装备的参数指标范围、预置的关键字和预设的预警阀值，通过自动或手动的方式，自动判断记录的装备状态是否存在异常、分析的结果中是否有超出指标范围，是否有与预置的关键字匹配的异常、是否有达到预警阀值的异常，从而判断装备的工作状态是否良好。如果发现故障和异常信息及时向用户进行健康预警，告知用户装备异常的发生时刻和对应的装备部件，方便用户及时掌握装备状态和故障部位，并自动产生异常告警报表。

2.2 系统数据库实现

为了满足系统对解析、清洗、规范化等数据存储的要求，设计了 15 个物理结构表来满足数据库要求，它们分别是用户表（t_user）、装备类型表（t_equipType）、适配器表（t_adapter）、指标库表（t_standardValues）、装备表（t_equip）、分系统表（t_subSystem）、部件表（t_component）、状态参数表（t_parameter）、日志文件表（t_logFile）、日志表（t_log）、日志类型表（t_logType）、任务项目表（t_tasks）、日志等级表（t_logLevel）、健康等级表（t_healthLevel）、健康评估标准表（t_evalutionStandard）。结合 mysql 语言的特性，运用 MySQL Workbench 建模软件，构建了如图 4 所示的数据库逻辑结构和 E－R 关系图。

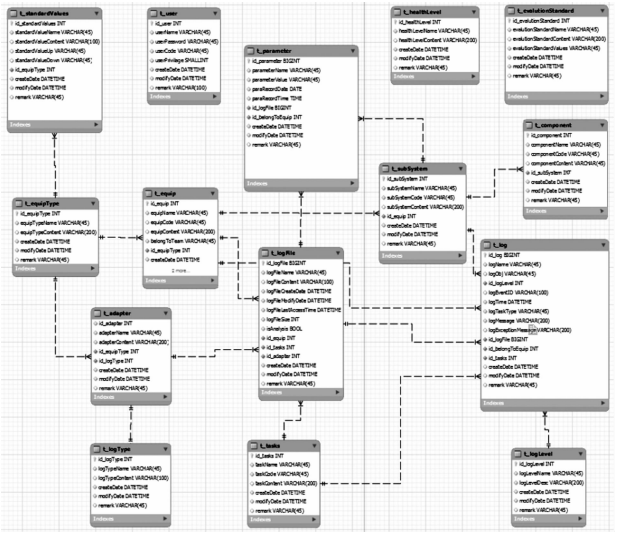

图 4 系统 E－R 图及逻辑结构设计

2.3 系统关键技术

该系统采用 C＃作为开发软件，MySQL 数据库作为数据管理软件。在实现的过程中，应用下列关键算法解决开发过程中遇到的关键问题。

1）利用正则表达式[7]解析原始日志，提高程序的灵活性和可配置性。

正则表达式是一种用于模式匹配和替换的强有力工具，在词法分析程序中大量使用，使用正则表达式可以大大简化字符串操作代码的编写。由于不同种类装备产生的日志文件格式不统一，同一装备的不同分系统日志格式也不统一，因此可以根据不同的日志需求设计不同的正则表达式来对日志进行分析。该系统采用正则表达式，目的是为了提高程序的灵活性、可配置性和自适应性。系统在使用过程中，可根据要解析日志的需求，从预存入数据库中的适配器表中读出正则表达式，然后对输入的日志根据该正则表达式进行解析，当有日志格式发生改变时，可以制定专用的正则表达式，使系统适应新的日志文件格式。

2）使用数据挖掘中的简单关联规则及序列关联规则，深入分析装备发生故障的序列规则和故障部位之间的关联关系。

使用关联规则从日志中发现彼此相关的事件。例如：如果装备的某个部件指示异常，可以从日志文件中搜索相关信息，来寻找任何与该异常事件相关的记录，这样可以对发生的异常事件进行综合分析。在进行关联规则分析过程中，最常用的就是 Agrawal 等人提出的 Apriori 算法[8]，过程包括两部分，首先是产生频繁项集，随后依据频繁项集产生关联规则。通过关联规则的挖掘，发现装备操作手的一些操作习惯，找出装备异常部位的关联关系，提高定位装备故障准确度。

3）结合云重心评估法[9]，构建装备健康状态评估模型[10]，采用加权偏离度来衡量装备的健康状态，实现装备健康状况的自动评估。

实际运行中的装备状态具有随机性，日志数据也有不确定性，日志的来源涵盖设备硬件、软件和运行的操作系统等方面，装备健康管理需要在它们之间建立映射和转换关系，通过确定健康等级评价集、建立评估对象的指标体系、确定各指标的权重、求各指标的云模型、用加权偏离度来衡量云重心的改变等步骤来形成定量可比的评价意见。云重心评估法能够实现定性属性值和定量属性值之间的转换，能够按照评估模型的要求评估装备的健康状况，故将此方法应用于该系统的装备健康状态评估之中，采用加权偏离度来衡量装备的健康状况，实现装备健康状况的自动评估，完成了装备健康状况的定量衡量。

3 实验结果与分析

3.1 日志文件的分析与处理

首先结合实际日志文件构建了正则表达式，实现了对日志信息的提取分析、快速精确查找以及匹配替换。

3.1.1 构建正则表达式

采用某型设备在某次工作过程中产生的“状态日志”文件来设计正则表达式，原始日志的格式如图 5 所示。

```
2017－07－01 07:29:30 FM Band A Input Power: ＊＊dbm
2017－07－01 07:29:30 Extended Band B: Not Captured
2017－07－01 07:29:30 Central Module: Error
2017－07－01 07:29:30 Central Case Power Supply: Normal
2017－07－01 07:30:35.4730 Result Receive Order: Set Primary & Secondary System: A as Primary, B as Secondary
2017－07－01 10:01:30 Central Component D13: Normal
2017－07－01 10:01:39 XXW Amplifier Output: Forbidden
2017－07－01 10:03:15 Transducer: Broken Chain
……
```

图 5 某型设备分系统 A 状态日志片段

分析日志信息可知，该文件包含了不同分系统、不同部件的多种信息，在构建正则表达式时，应分别进行检索和提取[11-12]。

针对图 5 所示的分系统 A，构建了两种正则表达式分别用来提取设备故障状态和参数数据：

1）以“Broken Chain | Error | Forbidden”为关键字，从“状态日志”中提取不同分系统工作状态为异常的正则表达式为：

\w* [Broken Chain | Error | Forbidden]；

2）从日志文件中提取不同分系统和不同部件设备参数数据的正则表达式为：

[']\d+[-]\d+[-]\d+)\s(\d+[:]\d+[:]\d+)\s(\w+)\s*(\w+[/]*\w+)|(\w+[/]*\w+)\s(\w*[-]*\w*[\.]*\w*[\+]*\w*[/]*\w+)\s+；

3.1.2　运用正则表达式处理日志文件。

随机选取一次装备运行事件，采用本文方法为设备分系统 A“状态日志”构建正则表达式“\w* [Broken Chain | Error | Forbidden]”，根据该表达式进行设备异常提取，结果如下：

2017－07－01 07:29:30 Central Module: Error

2017－07－01 10:01:39 XXW Amplifier Output: Forbidden

2017－07－01 10:03:15 Transducer: Broken Chain

……

3.2　故障关联分析与装备健康评估

选择 9 次设备运行事件，统计日志中的所有故障事件并编号如下：

I_1: FM Band A Input Power: Invalid

I_2: Extended Band C: Not Captured

I_3: XXW Amplifier Output: Forbidden

I_4: Secondary Module: Error

I_5: Central Case Power Supply: Anomaly

令设备故障集合 $\Omega=\{I_1, I_2, \cdots\cdots, I_m\}$，基于 Apriori 原理挖掘 Ω 的所有可能组合及其产生的频繁项集。首先对该实例中所有故障事件建立数据库，见表 1。

表 1　设备日志中的故障事件

设备日志	故障列表	设备日志	故障列表
Log001	I_1, I_3	Log006	I_1, I_2, I_4
Log002	I_1, I_2, I_3, I_5	Log007	I_1, I_2, I_5
Log003	I_1, I_2, I_3	Log008	I_2, I_4
Log004	I_1, I_3	Log009	I_2, I_3
Log005	I_2, I_3		

首先定义最小支持度阈值 min_support＝20%，最小置信度阈值 min_confidence＝60%。为找到频繁 K 项集，对该数据库进行逐层迭代搜索，通过第 1 次搜索找到频繁 1 项集并记为 L_1，第 2 次搜索找到频繁 2 项集 L_2……每找出每个 L_k 就进行了一次数据库的完整扫描，最终找到频繁 3 项集为 $\{I_1, I_2, I_3\}$ 和 $\{I_1, I_2, I_5\}$，挖掘其所有非空子集的关联规则，如表 2 所示。

表 2　故障关联分析

关联规则	置信度	关联规则	置信度
$\{I_1, I_2\}\Rightarrow I_3$	2/4＝50%	$\{I_1, I_2\}\Rightarrow I_5$	2/4＝50%
$\{I_1, I_3\}\Rightarrow I_2$	2/4＝50%	$\{I_1, I_5\}\Rightarrow I_2$	2/2＝100%
$\{I_2, I_3\}\Rightarrow I_1$	2/4＝50%	$\{I_2, I_5\}\Rightarrow I_1$	2/2＝100%
$I_1\Rightarrow\{I_2, I_3\}$	2/6＝33%	$I_1\Rightarrow\{I_2, I_5\}$	2/6＝33%
$I_2\Rightarrow\{I_1, I_3\}$	2/7＝29%	$I_2\Rightarrow\{I_1, I_5\}$	2/7＝29%
$I_3\Rightarrow\{I_1, I_2\}$	2/6＝33%	$I_5\Rightarrow\{I_1, I_2\}$	2/2＝100%

其中满足 min_support 的关联规则为 $\{I_1, I_5\}\Rightarrow I_2$，$\{I_2, I_5\}\Rightarrow I_1$ 和 $I_5\Rightarrow\{I_1, I_2\}$。在本装备健康管理系统中，健康预警评估界面如图 6、7 所示。

图 6　健康预警界面

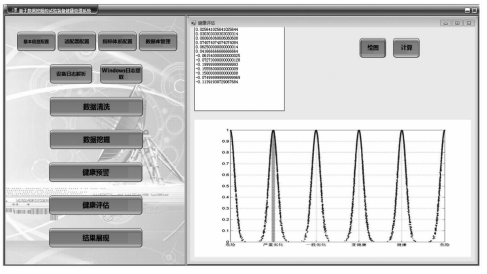

图 7　健康评估界面

4　结论

本文将数据挖掘的相关理论运用于装备的日志分析，设计实现了一个装备健康管理系统，该系统已在某型测控装备中得到应用。从实际应用情况来看，该系统能够加快装备日志自动化处理速度，提高分析结果的准确率和装备异常模块定位的准确性，提升整套装备的可维护性，实现了装备健康状态的自动评估和健康预警、装备履历日志的自动电子化和质量控制文档过程的电子化。

参考文献：

[1] 沈金明. 基于系统日志的计算机网络用户行为取证分析系统的研究与实现 [D]. 南京：东南大学，2006.

[2] 张如云. 基于日志文件的数据挖掘机理分析和研究 [J]. 微型机与应用，2015，34 (18).

[3] 路广勋，李建增，李鹏俊. 基于云重心评估法的发射场液压泵的健康状态评估 [J]. 计算机测量与控制，2014，22 (3)：0800－0802.

[4] 李烈彪，张海鹏，周亚峰. Web 日志挖掘中数据预处理方法的研究 [J]. 计算机技术与发展，2007 (7).

[5] 王永贵，林　琳，刘宪国. 结合双粒子群和 K－means 的混合文本聚类算法 [J]. 计算机应用研究，2014，31 (2).

[6] 尚永爽，许爱强，李文海. 航空装备视情维修的动态性研究 [J]. 装备指挥技术学院学报，2010，21 (6).

[7] 王　成，杨建华，蒋光伟. 正则表达式在测量数据处理中的应用 [J]. 测绘科学，2011，26 (2).

[8] 陈志飞，冯　钧. 一种基于 Apriori 算法的优化挖掘算法 [J]. 计算机与现代化，2016，235 (9).

[9] 齐伟伟，夏良华，李　敏，等. 基于云重心评估法的装备健康状态评估 [J]. 火力与指挥控制，2012，37 (4).

[10] 钟诗胜，谭治学. 雷达发射机健康状态评价技术研究 [J]. 现代雷达，2014，36 (6).

[11] 李　璋，杜慧敏，张丽果. 基于分布式存储的正则表达式匹配算法设计与实现 [J]. 计算机科学，2013，40 (3)：74－76，99.

[12] 邵　英，陆月明. 基于优化正则表达式的文本告警信息的提取与分析 [J]. 微型电脑应用，2010 (26).