文章编号:1671-4598(2017)08-0039-04 DOI:10.16526/j.cnki.11—4762/tp.2017.08.011 中图分类号:TN407 文献标识码:A

基于配置关联关系的信息系统误配置检测技术

向华伟，吕 森，张雪坚
（云南电网有限责任公司 信息中心，昆明 650217）

摘要：随着软件系统变得更加复杂和可配置，由于错误配置而导致的故障正成为关键问题；这种故障的诊断和修复需要跨越软件本身及其运行环境进行分析，使得其处理过程十分困难，且修理费用极高；为解决这种故障带来的较为严重的经济损失、安全隐患和功能故障；基于配置项之间隐含的关联关系及其运行环境，设计了基于信息系统配置关联关系的配置错误检测系统技术，利用给定的大量样本配置训练，形成配置项关联关系与检测规则，通过发掘信息系统各组件配置项之间的关联关系并利用这种关联进行配置项交叉检验，能够有效检测系统的错误配置；通过模拟测试表明，所提错误配置检出率达到了90%以上，在大型企业中具有广泛的应用前景，为未来优化信息误配置检测技术提供建设性方向方法。

关键词：配置关联；信息系统；检测规则；配置检测

Error Detection of Information System Based on Configuration Relation

Xiang Huawei, Lv Yao, Zhang Xuejian
(Yunnan Power Grid Corp. Ltd., Information Center, Kunming 650217, China)

Abstract: As software systems become more complex and configurable, failures due to faulty configuration are becoming critical issues. The fault diagnosis and repair need to be analyzed by the software itself and its operating environment, which makes the process very difficult and the cost of repair is very high. In order to solve the serious economic loss, safety hidden trouble and functional fault. The implied relationship between configuration items and its operating environment based on the design of the testing information system configuration configuration error correlation system based on the use of a large number of training sample configuration is given, the formation of configuration items associated with the detection rules, by exploring the information system of each component configuration relationship between items and use this configuration cross correlation inspection, can effectively detect system configuration errors. The simulation results show that the detection rate of the proposed error configuration is more than 90%, which has a wide application prospect in large enterprises, and provides a constructive way to optimize the information error detection technology in the future.

Keywords: configuration association; information system; detection rule; configuration detection

0 概述

随着软件系统功能日益丰富、应用更加灵活，其配置变得更加复杂。例如 MySQL 服务器和 Apache 服务器每个软件均含有 200 个以上的配置项[1-3]。这使得正确的配置软件系统已成为非常复杂的工作，并且容易出错。相关研究已表明配置错误已经非常常见，并且会对企业造成较大损失。配置错误不仅会引起系统不可用导致的业务中断，还会消耗大量资源进行故障排除[4-6]。例如某大型商业公司的客户支持数据库中，超过27%的故障单都与配置错误有关。同时，很多机构会按照最佳实践应用安全策略和性能策略，仅满足功能要求的配置设置多数无法满足相关安全及性能策略，这也会导致安全弱点或性能异常，检测这些次优配置也是非常必要的[7-9]。

1 相关研究现状

为了将配置错误导致的损失降低到最小，是在应用配置之前自动检查一组配置设置，提前发现潜在的设置错误，类似于源代码审计发现系统漏洞[10]。然而当前使用的大多数配置文件与编程语言相比缺乏丰富的结构和语义信息，难以直接进行复杂的错误分析[11-12]。为了克服该限制，一些研究通过收集分析大量配置项的常用值，形成训练集合，并将偏离常用值的配置内容标记为潜在的错误配置。

但该方法简单的将每个配置项作为独立的字符串进行处理，仅在某些情况下是有效的，其应用场景和效果均有限。例如，针对 PHP 系统来说[13]，其 extension _ dir 配置项用于指定扩展插件的目录，PHP 软件从该目录下搜索相关扩展插件。当其被配置为一个具体的文件，如“usr \ bin \ php \ php _ mysql”时，将会导致该 mysql 模块无法正常加载。又比如配置 mysql 数据存储目录时，其 datadir 指定了 mysql 数据的保存位置，如 datadir=/var/lib/mysql，但如果我们指定的目录 mysql 的启动账户没有该目录的访问权限，那么 mysql 将由于数据文件拒绝访问导致启动错误。

上述两个例子在实际检测时，仅通过对配置文件中的配置内容进行分析，均无法发现配置问题，其原因在于，现有检测方法仅分析配置项中的字符串，而这两个案例中的配置项的内容均与环境有关，此类配置内容随环境不同多种多样，难以形成异常配置模型。对于 PHP 的案例来说，检测系统不知道用户配置的“usr \ bin \ php \ php _ mysql”是一个目录还是一个具体的文件，对于 mysql 的案例来说，检测系统不知道启动 mysql 的账户是否有权对/var/lib/mysql 进行读写。

2 基于关联信息的配置错误检测系统设计

由于配置设置涉及应用本身和应用的操作环境，为了解决现有配置正确性检测方法仅针对配置内容本身进行检测的缺

收稿日期:2017-04-20； 修回日期:2017-05-11。

作者简介:向华伟(1984-),男,云南玉溪人,大学,高级工程师,主要从事信息系统运维、自动化运维方向的研究。

陷，需要将配置分析的范围从单个配置项扩展到多个配置文件以及操作系统相关环境设置。

2.1 配置项关联关系梳理

对于一个完整的信息系统，操作系统、中间件、数据库、网络安全组件等软件均为相互关联的关系，各组件之间的关联关系体现在环境信息上[14-15]。操作系统安装于服务器硬件，网络配置安装于网络设备，并且网络配置依赖于物理拓扑连接[16]。而数据库、中间件则需要安装于操作系统之上，并且数据库和中间件之间为关联关系；操作系统、数据库、中间件又与网络配置相关联。图中标识为安装的关系，也就包含了相关环境的个性化配置。如图 1 所示。

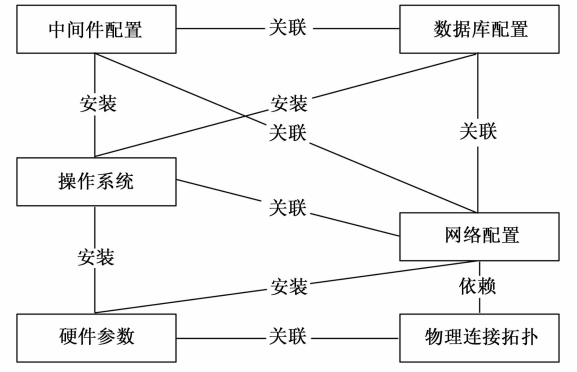

图 1 信息系统各软件关联关系

而对于单个软件的配置层面，操作系统安全配置依赖组件安装，数据库、中间件安全配置依赖于操作系统安全配置。中间件功能与数据库功能配置关联，如图 2 所示。

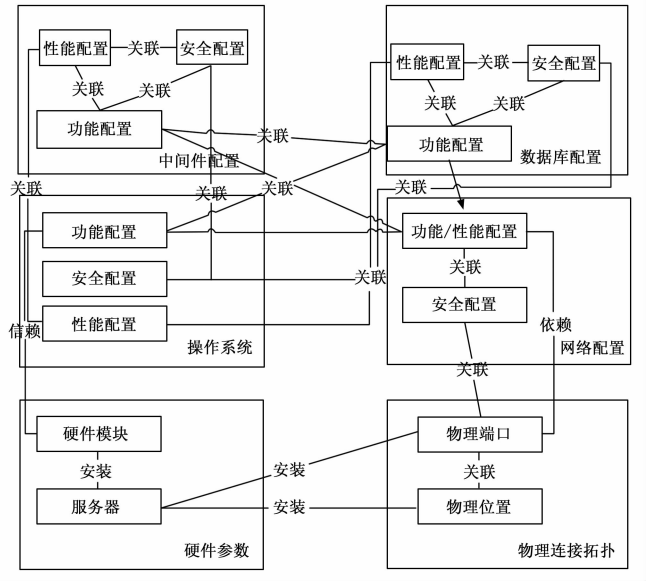

图 2 软件配置项间的关联关系

2.2 检测系统架构

为实现配置错误检测，需要自动提取各软件、操作系统、网络设备的配置信息，将配置解析后应用数据挖掘算法实现配置项的关联关系自动检测，根据关联关系推断出配置规则模板，配置检测模块根据相关规则对配置进行检查以发现配置异常。系统架构如图 3 所示。

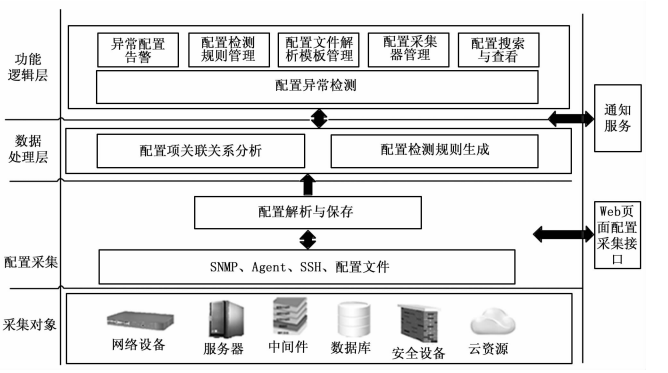

图 3 检测系统架构图

1）配置采集层：通过客户机代理、ssh、配置文件、SNMP 等方式采集原始的配置信息并保存；同时将原始配置信息应用配置解析模板进行解析，转换为 key－value 格式的配置记录，保存到配置库中。采集工作完成后，整个信息系统相关的配置信息均已统一保存，可被数据处理层和功能逻辑层调用。

2）数据处理层：应用优化的关联规则挖掘算法（Apriori 和 FP－Growth）挖掘配置项中的关联关系，使用前首先利用现有的正常运行的系统配置训练关联模型，在形成关联关系后供配置检测规则生成模块使用。规则生成模块使用关联关系和数据集训练生成关联规则，通过应用规则模板简化规则生成的难度。

3）功能逻辑层：功能逻辑层根据配置库和配置检测规则进行配置异常检测，当实际配置与检测规则不一致时进行告警。同时提供自定义配置规则的管理、配置文件解析模板管理、配置采集器管理。配备配置搜索与查询，与配置告警关联，可以之间从告警信息跳转到对应的原始配置文件。

2.3 配置项关联关系分析

一般而言，配置项关联首先表现在配置项的类型是一致的，因此关联关系的生成需要根据配置项的类型进行匹配，例如对于 mysql 服务器的监听 IP 和端口设置，其类型为 IP 和端口，而操作系统、防火墙、中间件相关的配置项也一定会是 IP 和端口，操作系统的网络接口 IP、防火墙开放的端口、中间件连接数据库文件中的数据库 IP 和端口均是相关联的，应用该思路能够极大的简化关联关系生成时的时间复杂度和空间复杂度，也能够消除一些不相关的噪声影响。

2.3.1 类型匹配与学习

我们设计了一系列类型及其匹配的正则表达式用于单个配置项类型识别，当配置项的值完全匹配时，即可初步确认该配置项类型，随后应用相关命令验证该类型的准确。例如当分析器检测到某配置项的值为“/var/www/abc”时，首先将其初步分类为“路径”，随后调用路径类型相关命令检测该路径是相对路径还是绝对路径，以确认路径的存在。当此操作确认后，将此配置项类型根据实际分类为“相对路径”或“绝对路径”。

当然某些类型难以被准确验证，可以通过后续的训练，统计大量样本中配置项的内容，应用分类算法得出该配置项的最终确认类型。

2.3.2 关联关系生成

通过为每个配置项自动设置了类型，相关的关联关系可以

表 1　部分典型配置项类型模板

配置项类型	匹配规则	额外检验方法
绝对路径	/.+(/.+)*	Ls 命令
相对路径	/?.+/(.+)*	Ls 命令、pwd 命令、读取环境变量
用户名	[a-zA-Z][a-zA-Z0-9]*	检查/etc/passwd
用户组	[a-zA-Z][a-zA-Z0-9]*	检查/etc/group
IP 地址	[\d]{1,3}(.[\d]{1,3}){3}	检查本机网络接口 ip、ping
端口	[\d]+	/etc/services、netstat、ping
文件名	[\w-]+.[\w-]+	Ls 命令
URL	[a-z]+://.*	Curl
字符集	[\w]+	国际标准
语言	[a-zA-Z]{2}	国际标准
空间大小	[\d]+[KMGT]	无额外验证
布尔值	Values Set	无额外验证
字符串	无	无额外验证
数值	[0-9]+[.0-9]*	无额外验证

首先基于类型进行关联，对于那些通用的配置项类型如字符串、数值等类型可进一步应用关联算法训练生成。

1）基于类型的关联关系生成：仍然以 Mysql 的监听地址为例。经过上述步骤处理，我们已经对整个信息系统的日志完成了采集、解析及类型匹配。此时 mysql 配置文件中的监听地址已经被识别为 ip 地址，那么相应的中间件的数据库连接文件中的数据库地址、操作系统网络接口的地址均被识别为 IP 地址，而这时这些 IP 地址类型的配置项就产生了关联关系，相应的检测规则也可以据此生成。

2）基于关联算法的关联关系生成：由于有特定含义的类型数量有限，大量的配置项均为通用的类型，同时类型相同的配置项也并非总是存在关联关系，例如 mysql 的监听地址与本机 vpn 连接使用的远端服务器地址类型相同，但并没有存在关联关系。此时需要将整理好的运行正常、配置已调优的信息系统各组件的配置形成训练数据，关联算法即可根据配置值总是关联出现的频率自动形成配置项的关联关系。

2.4　检测规则生成

根据 3.3 节描述的方法生成配置项间的关联关系后，可以进行检测规则的生成工作。由于配置项的具体内容与实际应用环境紧密相连，单纯依靠数据挖掘工具形成的检测规则效率低下，在大型信息系统环境下甚至不可行。本文根据运维经验设计了一系列检测规则模板，用于指导数据挖掘系统形成更加有效的规则。

例如，根据前文描述的 mysql 数据保存路径的场景，设计规则模板为 [<组件>_<路径>] grant [<组件>_<用户名>]，其中尖括号中的为占位符。该规则在生成时将首先枚举全部的组件和配置项，比如本例中产生的一条有效规则为：mysql_datadir grant centos_mysql，其含义为 centos 的账户 mysql 必须拥有访问 mysql 的 datadir 目录的权限。

本文通过人工分析相关配置项的关联关系，预先设定了部分规则生成模板，同时系统设计了检测规则管理，使用者可以自行定义、添加、删除相关模板。

表 2　部分预置的规则生成模板

模板	说明
[<组件>_<类型>] === [<组件>_<类型>]	一个配置项应与另外配置项相关联，其值相等，并且类型一致
[<组件>_<IP>] =[<组件>_<IP>]	一个 ip 地址配置项应与另外一个 ip 地址配置项的值属于同一子网
[<组件>_<路径>] grant [<组件>_<用户名>]	用户名配置项的值应有权访问文件路径配置项的值
[<组件>_<路径>] denied [<组件>_<用户名>]	用户名配置项的值无权访问文件路径配置项的值
[<组件>_<路径>] +[<组件>_<文件名>]=[<路径>]	某个组件的路径和某个组件的文件名拼接后的类型应该为路径
[<组件>_<字符串>] belong [<组件>_<字符串>]	一个字符串配置项的值应是另一个字符串配置项的子集
[<组件>_<用户名>] belong [<组件>_<用户组>]	一个用户名配置项的值应属于一个用户组配置项
[<组件>_<数值>] < [<组件>_<数值>]	一个数值配置项的值应小于另一个数值配置项
[<组件>_<空间大小>] < [<组件>_<空间大小>]	一个空间大小配置项的值应小于另一个空间大小配置项

对于每个模板系统将自动根据占位符使用所有匹配占位符的数据进行填充，形成大量的初始规则，由于规则生成是无状态的，可以并行对多个规则模板进行规则生成。

初始规则生成后，其中包含了大量的垃圾规则，其规则并无实际意义，这时需要对初始规则进行筛选，去除无效的规则。应用关联规则检测中常用的支持度和置信度可以进行此类筛选。支持度表征规则相关的具体配置项值出现的频率，置信度表征该规则的有效程度。由于某些配置项在不同系统、场景中的值通常相同。这些不经常改变的值对于规则检测来说没有意义，因此可以使用信息熵的概念去除变化较少的配置项相关的规则。

2.5　异常检测

当检测规则生成后，检测系统通过以下方法检测潜在的配置错误：

1）配置项名称错误：系统应用内置的配置项清单对比是否有在清单以外的相似的配置项，这很有可能是拼写错误；

2）关联错误：系统根据训练形成的检测规则检测不匹配的关联配置项，根据规则表达式计算实际值与预期值比较，报告出现的不一致情况作为配置异常告警；

3）配置值类型错误：系统读取每个配置项的当前值，识别该值的类型，并与训练生成的配置项类型对比、检验，当类型不一致或检验出错时进行告警；

4）可疑配置值检测：系统读取每个配置项的值，并于训练集中的对应配置项值进行对比，如果当前配置值在训练集中从未出现过，该配置项将作为可疑值进行提示。当出现多条配置项的值为可疑值时，按训练数据中对应配置项的值变化的频率从低到高进行配置项排序，按配置项顺序从高到低设置可疑等级。

3 实验结果分析

3.1 实验环境设置

为了验证改进误配置检测技术在信息系统上使用方面的有效性及可行性，需进行实验对比分析。实验采用 WIN7 系统，CPU 为 P4 1.7 G、内存为 512 M，其测试环境网络拓扑如图 4 所示。

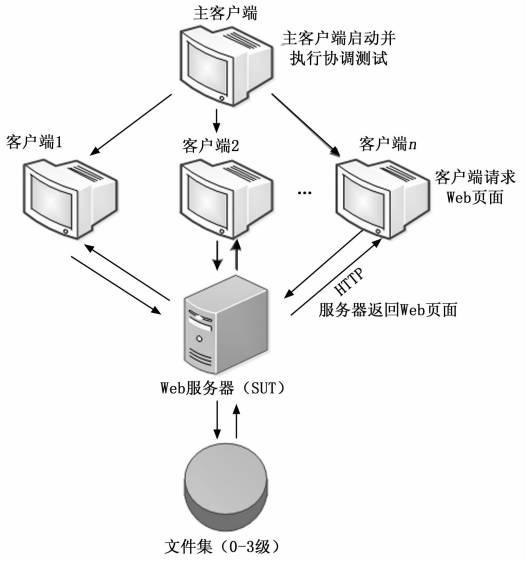

图 4 测试环境网络拓扑图

3.2 实验结果对比

为了证明本文提出的基于信息系统关联关系的误配置检测技术改进方法的有效性，将其进行采用导出策略法与改进误配置检测技术做对比分析，以错误配置检出率为指标进行实验分析，结果如图 5 所示。

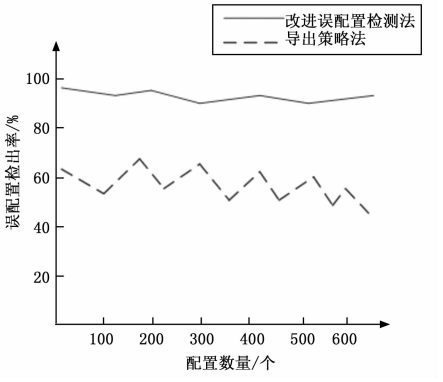

图 5 检测检出率对比分析

由图 5 可知，在检测数量相同的情况下，改进误配置检测法检出率明显高于导出策略法，改进误配置检测法误配置检出率平均在 90%，而导出策略法误配置检出率平均在 60%。而且，随着检测配置数量的不断增加，导出策略法误配置检出率也随之降低且波动幅度大，但相比改进误配置检测法，误配置检出率平缓稳定，不随配置数量的增多而产生波动。得出结论改进误配置检测法在误配置检出率方面性能明显优于导出策略法。这种检测方法优越之处在于其先通过为每个配置项自动设置了类型，有关的关联关系可以基于类型进行关联，对于那些通用的配置项类型可进一步应用关联算法训练生成，然后有效利用配置项之间的关联关系，从而完成检测规则的生成工作。由于配置项的具体内容与实际操作环境息息相关，所以应用改进误配置检测法才能够准确的发现配置错误，为系统的使用提供了更安全，更有效的操作环境，具有较好的应用价值。

4 总结

根据本文提出方法实现了检测系统原型，并基于原型对检测效率进行了测试，通过对 Apache、Mysql、SSH 服务进行了模拟测试。

在模拟测试中，选择不在训练数据中的 Apache、Mysql、SSH 服务配置文件，随机插入 15 个错误配置项，应用检测系统原型进行检测，Apache 检测到 14 个配置错误、Mysql 检测到 15 个错误、SSH 服务检测出 15 个错误。

测试表明，训练完成的配置错误检测系统能够利用配置项之间的关联关系较为准确的发现配置错误，在大型企业的信息化环境中进行现有配置检查、用户输入配置前的验证以及自动化运维系统配置修改脚本的检查中具有非常重要的意义，能够避免配置输入错误导致的非预期停机，极大的降低了故障处理和系统变更评审所需的人力资源，具有广泛的应用前景。

参考文献：

[1] 李 健. 路由优化的方式——策略路由及其配置方法 [J]. 电子技术与软件工程，2016 (12)：25－25.
[2] 舒 鹏，王 松. 基于系统描述自动生成二次设备关联配置的方法 [J]. 浙江电力，2016，35 (7)：12－15.
[3] 林雪峰，宋跃忠，程 敏，等. 一种网元设备误配置检测方法及检测设备，CN105847065A [P]. 2016.
[4] 王雅青，徐 斌，谭国平. 基于 D2D 与 WLAN 共享的分布式频谱配置方案研究 [J]. 电子设计工程，2016，24 (14)：87－90.
[5] 刘海涛. 网络入侵检测算法研究 [J]. 电脑知识与技术，2015 (2X)：42－43.
[6] 李 勒，洪爽俊，张 驰. 基于部分配置信息的错误数据注入攻击 [J]. 微电子学与计算机，2014 (2)：85－89.
[7] 蒋建春，陈慧玲，邓 露，等. 基于多核实时操作系统的配置工具设计 [J]. 计算机应用，2016，36 (3)：765－769.
[8] 舒 鹏，王 松. 基于系统描述自动生成二次设备关联配置的方法 [J]. 浙江电力，2016，35 (7)：12－15.
[9] 肖 宁. 基于高分辨一维距离像的雷达自动目标识别技术研究 [D]. 西安：西安电子科技大学，2014.
[10] 孙 莉. 电子文件管理系统的架构、核心算法及其实现 [J]. 图书馆理论与实践，2015 (11)：67－69.
[11] 温树峰，孙 丹，王珍珍. 智能变电站配置文件错误分析 [J]. 能源与节能，2015 (1)：176－178.
[12] 李晓宁. 电子化文件站配置探索 [J]. 电子技术与软件工程，2015 (9)：76－76.
[13] 马 静. 基于 PHP 的高校图片管理系统的设计与实现 [J]. 自动化与仪器仪表，2015 (4)：126－127.
[14] 汪 溢，马 凯，黄 曙，等. 变电站二次设备参数配置方法和系统，CN104537572A [P]. 2015.
[15] 黄 曙，马 凯，汪 溢，等. 智能变电站二次设备运行参数配置方法，CN104135069A [P]. 2014.
[16] 李伦红. 网络结构物理联接拓扑发现应用研究 [D]. 西安：长安大学，2015.