doi:10.11857/j.issn.1674-5124.2014.04.025

工业控制系统软件的用户操作异常检测方法

文元美，余雪晨
(广东工业大学信息工程学院,广东 广州 510006)

摘 要:针对工业控制网络终端控制系统软件的用户行为安全问题,提出用户操作层面行为可信评估的方法。首先从软件日志文档中提取出整数数列的历史序列,然后引入 Apriori 方法进行用户操作行为序列频繁模式挖掘,并基于挖掘出的模式集建立概率矩阵，最后通过提取当前用户行为序列，基于 BLAST-SSAHA 算法与历史序列模式集序列匹配,对用户行为可信性进行检测,为后续系统平台可信性评估提供依据。实验表明 文章提出的方法是有效可行的。

关键词:用户操作异常 数据挖掘 模式匹配 工业控制网
中图分类号:TP301;TP311.53;TP206+.3;O224 文献标志码:A 文章编号:1674-5124(2014)04-0098-04

User operation anomaly detection method of software in industrial control system

WEN Yuan-mei, YU Xue-chen
(School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: A behavior trust evaluation method was proposed to solve the problem of user behavior security of terminal control system software for industrial control network. Firstly, history sequences in the form of numerical array were extracted from the log file of software. Then, the Apriori mining algorithm was exploited to obtain the frequent sequential patterns from user action sequence, and a matrix of probabilities was established upon mined patterns. Finally, the credibility of user behavior was detected by extracting the current user behavior sequence to match with the sequence in historical sequence pattern set based on BLAST-SSAHA algorithm. These results can provide the basis for evaluating the credibility of the system. The experimental and analytical results show that the effectiveness and feasibility of the methods are validated.

Key words: user abnormal operation; data mining; pattern matching; inclustrial control netnork

0 引 言

随着信息技术的发展,工业控制系统的安全性逐渐成为人们关注的热点。美国国家标准与技术研究院发布《ICS Security》(2011 年),制定了工业控制系统(industry control system,ICS)安全管理规范。

Youngjoon Won 等[1](2012)研究了基于 IP 的无线控制系统与网络的故障检测方法,分析传统以太网/IP 网络故障诊断方法及其局限性,改进对 ICN 故障独特的交通特性及分类,同时提出了一种故障诊断、预测及自适应决策方法,并利用来自 POSCO 钢铁冶炼公司的实际数据对其进行验证;美国伯克利大学的 Roosta,T. Nilsson[2](2008)研究了一种基于模型的无线过程控制系统入侵检测系统(intrusion detection system,IDS),IDS 模仿无线传感网络常态行为实体与检测攻击,该模型可用于检测未知攻击;芬兰技术研究中心[3](2011)在 MOVERTI(monitoring for network security status in modern data networks)项目报告中,分析了网络安全监控系统中一些特定操作的环境威胁,通过演化网络数据流特征得出安全设备精确报警阈值等,监测系统整体设备区域安全状态并提交给操作者。

收稿日期:2013-12-09;收到修改稿日期:2014-01-24
作者简介:文元美(1968-),女,湖北荆州市人,副教授,博士,主要从事智能信息处理与可信研究。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

当前工业控制系统网络入侵检测主要是通过对整体设备网络进行监控，建立设备间通信流实体模型。对模型进行训练，分析入侵检测算法，提取入侵特征值，从而检测网络入侵。同时利用预测算法基于模型实例来抵御未知入侵。但是，现在大部分模型都是针对特定工业控制网络，依据不同网络特征对入侵进行判定。

本文针对软件运行过程，从用户操作层面提出一种用户操作序列的工业控制系统终端软件异常检测方法，给出了整体设计架构与方案。

1　用户操作层面行为可信架构与设计方案

一位可信用户对软件的操作行为在一段时间内会维持某种固定的特征，而非此用户本人的操作习惯会发生一定随机性的改变，因此本文从操作层面提出用户操作行为频繁序列，与当前用户操作序列匹配，以期提高物联网环境下终端软件的使用安全性[4]。

用户操作层面行为可信包括用户频繁序列获取与当前用户行为序列模式匹配两部分，用户操作层面行为可信总体功能架构如图 1 所示。

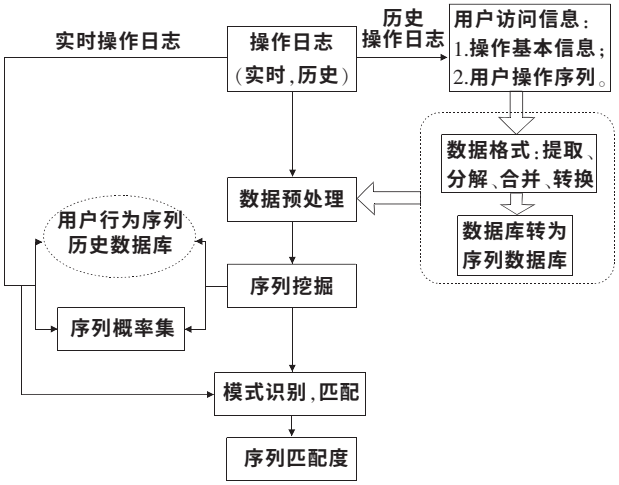

图 1　用户操作层面行为可信总体功能架构

用户对软件的操作行为信息通过操作日志获取，对获取的日志信息进行数据预处理，包括数据格式提取、分解、合并及转换[5]。依据行为信息时间戳形成有序操作序列，生成序列数据库。对于序列数据库中的信息，通过模式挖掘算法进行序列挖掘，结果形成用户行为序列历史数据库及序列概率集。

对运行时监测收集的用户操作日志文档进行预处理，获得当前用户序列，与历史数据库中的用户行为序列进行模式识别与匹配，得到序列匹配度。

工业系统软件用户行为操作模式异常检测方案如图 2 所示。

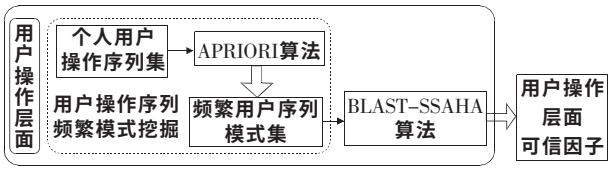

图 2　用户操作层面行为可信设计方案

在用户操作序列频繁模式挖掘中，本文采用 Apriori 算法来对个人序列数据库进行序列挖掘[6]。在模式匹配过程中，采用 BLAST-SSAHA 算法，将当前用户序列与历史数据库中的正常操作序列进行比对，判断当前用户序列的相似度[7]。当相似度低于某一阈值，则认为此段操作异常。

2　用户操作异常检测关键算法

2.1　用户行为序列的生成

当用户登录时，对于用户的每一个 button 的操作事件，后台都会向日志数据库表中添加一条记录，记录用户的操作时间、操作类型、操作类别编号等。

对原始日志文件中的数据进行提取、分解、合并、最后转换为适合进行数据挖掘的数据格式。对数据的清理使用建立规则库的方法，在程序中读入过滤规则进行数据清理，得到的数据重新保存到数据库表中，等待进一步处理[8]。

通过日志时间戳，查找数据库表中数据，依时间段形成用户操作序列，将形成的操作序列存储至序列数据库中。

2.2　用户行为序列模式挖掘

本文采用经典的频繁序列模式挖掘算法 Apriori 对用户行为序列进行挖掘，生成序列模式。Apriori 算法首先基于预定的最小支持度找出所有的频集，然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。一旦规则生成，那么只有那些大于用户给定的最小可信度的规则才被留下来，生成所有频集[9]。

定义 1　设 $I=\{I_1,I_2,\cdots,I_m\}$ 是项的集合，包含 k 个项的集合称为 k 项集。设任务相关的数据 D 是数据库事务的集合。

定义 2　设 A、B 分别是一个项集，其中 $A\subset I$，$B\subset I$，并且 $A\cap B=\Phi$。A、B 关联，表示为 $A\Rightarrow B$。规则 $A\Rightarrow B$ 在事物集 D 中成立，具有支持度 s，s 是 D 中包含 $P(A\cup B)$ 的百分比。

$$\text{support}\ (A\Rightarrow B)=P(A\cup B) \tag{1}$$

其中 $P(A\cup B)$ 表示事务包含集合 A 和 B 的并（即包含 A 和 B 中的每个项）的概率。

给定一个最小支持度 min_sup，频繁序列模式的集合包含所有支持度不小于 min_sup 的序列。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

性质　频繁项集的所有非空子集也必须是频繁的。如果项集 I 不满足最小支持度阈值 min_sup，则 I 不是频繁的，即 $P(I)$ <min_sup。如果项 A 添加到项集 I，则结果项集（即 $I \cup A$）不可能比 I 更频繁出现。因此，$I \subset A$ 也不是频繁的，即 $P(I \cup A)$ <min_sup。

定义 3　将 L_{k-1} 与自身连接产生候选 k 项集的集合。该候选项集合记作 C_k。C_k 是 L_k 的超集，C_k 的成员可以是也可以不是频繁的，但所有的频繁 k 项集都包含在 C_k 中。C_k 的形成包括连接与剪枝两部分。

⑴ 连接。设 l_1 和 l_2 是 L_{k-1} 中的项集。记号 $l_i[j]$ 表示 l_i 中的第 j 项。为方便起见，Apriori 假定事物或项集中的项按字典次序排序。对于（$k-1$）项集 l_i，意味将项排序，使 $l_i[1]<l_i[2]<\cdots<l_i[k-1]$。执行连接 $L_{k-1} \triangleright\triangleleft L_{k-1}$，产生项集 C_k。

⑵ 剪枝。扫描数据库，确定 C_k 中每个候选的计数，从而确定 L_k（即根据定义，计数值不小于最小支持度计数的所有候选是频繁的，从而属于 L_k）。然而，C_k 可能很大，这样所涉及的计算量就很大。可以使用 Apriori 性质，如果候选 k 项集的（$k-1$）项子集不在其中，则该候选也不可能是频繁的，从而可以从 C_k 中删除。

用户行为序列模式挖掘步骤包括：

⑴ 通过扫描数据库，累积每个项的计数，并收集满足最小支持度的项，找出频繁 1 项集的集合。该集合记做 L_1。

⑵ L_1 用于寻找频繁 2 项集的集合 L_2，L_2 用于找 L_3，如此下去，直到不能再找到频繁 k 项集。

2.3　当前用户异常行为操作模式匹配

以最近的用户操作序列作为查询序列，与数据库中保存的该用户频繁操作序列进行比对，若查询序列与常用序列相似度较小，则可认为用户当前操作行为异常。

给定一个长度为 n 的序列 $S=<s_1,s_2,\cdots,s_n>$，S 中任意连续的 k 个元素构成 S 的一个 k 元组。则 S 中 k 元组的个数为 $n-k+1$，每个元组被赋予一定的权重。

定义 4　在序列 S 中，某个 k 元组开始的位置称为元组偏移。

定义 5　序列 S 中每个位置可能的取值个数称为 S 的序列基数。

用从 0 到 Sequence Base−1 的整数来表示 S 中所有可能出现的元素，若序列基数为 ω，一个 k 元组$<t_1,t_2,\cdots,t_k>$的权重 W 记为

$$W=\sum_{i=1}^{k}\omega^{k-1}\times t_i \qquad (2)$$

假定数据库 M 中有若干个序列，SSAHA 算法将每一个长度为 n 的序列划分成 $n-k+1$ 个 k 元组，生成一张 k 元组表 KT。KT 中的每一条记录对应 M 中的一个 k 元组，并由以下元素组成：<元组权重，所在序列编号，元组偏移>。其中元组权重用来在 KT 上建立一个聚类索引，由 W 的计算可知，对于给定的 k 和 ω，一共有 ω^k 种权重，每种权重对应一种 k 元组而通过所在序列编号和元组偏移就可以确定当前的元组在 M 中的位置。

得到 KT 之后，对于一个查询序列，现在就可以通过 BLAST 算法将其与 M 中的序列进行比对，具体过程为：

⑴ 查询序列也划分为若干个 k 元组。

⑵ 依次将查询序列中的 k 元组与 KT 进行比较，若 KT 中也存在一个相同的 k 元组，则称为发生了一次命中，记录下该元组在 M 中的位置，即其所在序列编号和元组偏移。若某个 k 元组在 KT 中不存在，则称为发生了一次偏离。

⑶ 利用⑵中的记录，将 M 中的序列按照其包含查询序列不同 k 元组的个数进行排序，可以看出，包含查询序列中不同 k 元组越多的序列，与查询序列的相似程度就越大。

通过一定的阈值，在 M 中选取前 n 个与查询序列最相似的序列，将查询序列与 M 中序列发生命中的片段双向扩展，进一步比较查询序列与该序列的相似程度，并最终确定与查询序列最为相似的序列。

3　测量实验与结果

3.1　用户操作序列模式挖掘实验

基于布尔关联规则的频繁序列挖掘算法 Apriori，使用逐层迭代方法通过候选产生找出频繁项集。向挖掘程序输入用户操作行为序列事务数据库表 AprioriDb 中数据，设定最小支持度 min_sup，程序运行后输出数据存储至事务数据库表 TEMP 中，该表中数据即为用户历史操作行为序列频繁项集。

用户操作序列频繁模式挖掘主要方法函数：

Apriori_Gen（int [,] L_k_1,int [,] C_k,string[] strItemSet,int k_1）：连接和剪枝。由 L_{k-1} 得到 C_k 候选集。

SearchSubItem（int[,] L_k_1,int[]TEMP,int k_1）：查找 C_k 中某一项的子集是否在 L_{k-1} 项集中，TEMP 数组用来存储子集。

C_k_Count_Sup（int[,] ITEM，int[,] C_K,int[] Count,int k_1）统计候选项集中每一项的支持度。

基于 Apriori 算法的频繁序列挖掘结果如图 3 所示，其中 Apriori 算法参数最小支持度设置为 0.2。

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

挖掘出的频繁项集 B={item,count},item 表示项集，count 表示对应项集支持度计数。观察结果可看出，满足当前最小支持度的最长序列长度为 4,{4,4,6,2} 是其中一条用户操作频繁序列。找出所有的用户行为序列频繁模式之后，将数据存入事务数据库表 TEMP,算法终止。

图 3　用户操作序列频繁模式挖掘实验结果

3.2　当前用户操作行为序列匹配实验

将当前用户操作序列作为输入提交给 BLAST 算法，当前用户操作序列长度为 N,BLAST 算法在数据库表 TEMP 中查找相应的 NKT 进行比对。在此，简单定义出现一次命中的得分为 α,出现一次偏离的得分为 β。假定当前序列中出现了 T 次命中,那么其与用户此次操作序列的相似度 Y 为

$$Y=T\times\alpha-(N-T)\times\beta \tag{3}$$

程序后台设置相似度 $Y\geqslant 0$ 时判定操作安全，$\alpha+\beta=1$。规定操作序列命中率 $d\geqslant 60\%$时判定操作安全,则 $\alpha=0.4$,$\beta=0.6$。

测试过程中,输入实验序列 S,序列长度 $C_n=\{6,3,9,7,4\}$,后台程序内部输出相似度 $Y_n=\{0.4,1.2,-1.4,-0.2,0.6\}$，从而可直接对不满足要求的序列进行判定,系统强制禁止用户下一步访问请求。

4　结束语

本文提出了针对物联网模式下客户终端软件用户操作行为序列的异常行为检测方法。该方法主要针对工业控制网络中需提供管理员权限的软件,通过数据挖掘在操作层面分别对当前操作行为进行分析,通过当前行为序列与历史行为序列集的相似度来判断是否出现了异常。通过分析可以看出,引入数据挖掘的方法可有效提高用户行为可信的识别效率。

在接下来的工作中,将在方案中增添异常项集数据库,并对用户序列模式匹配过程进行优化,提高判断的实时性,从而更有效判断用户行为的可信性。

参考文献

[1] Won Y J, Choi M J, Park B, et al. An approach for failure recognition in IP-based industrial control networks and systems [J]. International Journal of Network Management,2012,22(6):477-493.

[2] Roosta T, Nilsson D, Lindqvist U, et al. An intrusion detection system for wireless process control system[C]// America,2008.

[3] Ahonen P. Constructing network security monitoring systems[Z]. Vtt Research Notes,2011.

[4] 柴洪峰,李锐,王兴建,等. 基于数据挖掘的异常交易检测方法[J]. 计算机应用与软件,2013,30(1):165-171.

[5] 彭成,杨路明,满君丰. 网络化软件交互行为动态建模[J]. 电子学报,2013,41(2):314-320.

[6] 陈岭,陈元中,陈根才. 基于操作序列挖掘的 OLAP 查询推荐方法[J]. 东南大学学报,2011,41(3):499-503.

[7] Kundu A, Panigrahi S, Sural S, et al. BLAST-SSAHA hybridization for credit card fraud detection [J]. IEEE Transactions on Dependable and Secure Computing, 2009,6(4):309-315.

[8] 毛伊敏. 数据量频繁模式挖掘关键算法及其应用研究[D]. 长沙:中南大学,2011.

[9] Han J W, Kamber M, Pei J A. 数据挖掘:概念与技术[M]. 北京:机械工业出版社,2012:147-154.

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net