第 12卷 第 4期
2005年 8月

湖南商学院学报 (双月刊)
JOURNAL OF HUNAN BUSINESS COLLEGE

Vol. 12 No. 4
Aug. 2005

数据挖掘技术在安全审计中的应用

石 彪[1] 胡华平[2] 刘利枚[2]

(1. 湖南商学院 计算机与电子工程系, 长沙 410205;
2. 国防科技大学 计算机学院, 长沙 410073)

摘 要〕本文主要讨论如何通过具有学习能力的数据挖掘技术, 来实现网络日志的综合分析与智能安全审计。在介绍数据挖掘技术和常用安全审计方法的基础上, 给出了一种基于数据挖掘技术实现的安全审计引擎框架。

关键词〕数据挖掘; 安全审计; 网络日志; KDD Weka

中图分类号〕F239.1 **文献标识码**〕A **文章编号**〕1008-2107 (2005) 04-0084-03

一、数据挖掘概述

随着数据库技术的迅速发展以及数据库管理系统的广泛应用, 人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息, 人们希望能够对其进行更高层次的分析, 以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能, 但无法发现数据中存在的关系和规则, 无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段, 导致了"数据爆炸但知识贫乏"的现象。同时, 计算机技术的另一领域——人工智能自 1956年诞生之后取得了重大进展。经历了博弈时期、自然语言理解、知识工程等阶段, 目前的研究热点是机器学习。机器学习是用计算机模拟人类学习的一门科学, 比较成熟的算法有神经网络、遗传算法等。

用数据库管理系统来存储数据, 用机器学习的方法来分析数据, 挖掘大量数据背后的知识, 这两者的结合促成了数据库中的知识发现 (KDD: Knowledge Discovery in Databases) 的产生。数据挖掘是 KDD最核心的部分, 是采用机器学习、统计等方法进行知识学习的阶段。预测和描述是数据挖掘的主要任务。预测是指用一些变量或数据库中的若干字段预测其他感兴趣的变量或字段的值; 描述是指挖掘出数据库的一般特性。许多人将数据挖掘看成是数据库中的知识发现 (Knowledge Discovery in Database KDD) 的一部分, 这是狭义上的数据挖掘; 从广义的观点来看, 数据挖掘系统代表了 KDD的整个过程。KDD的目标是从大型数据集中获取有用知识, 它是一个交互式的半自动分析工具, 系统的用户应当对有关领域具备良好的理解力。KDD过程如图 1所示。

1. 确定发现任务的应用领域、背景知识和性质。

2. 准备相关的数据子集: 将分布在各处以各种形式存放的数据, 按照 KDD的需求收集过来, 并根据分析需求, 选择适当的和典型的数据, 缩小处理范围。

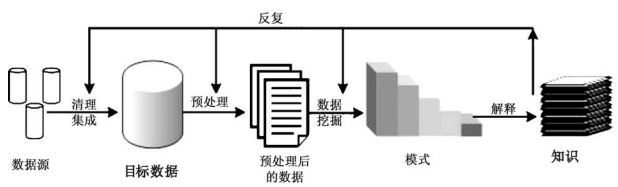

图 1 KDD过程示意图

3. 对数据进行预处理: 通过汇总或聚集操作将数据变换统一成适合挖掘的形式。

4. 进行数据挖掘, 发现模式并表达成易于理解的规则或树的形式: 模式是数据的一个子集的抽象表示, 它可以以人工的方式或自动的方式建立。

5. 评价和解释发现的模式: 根据设定目标 (通常为兴趣度度量), 利用专业知识, 对数据挖掘结果进行评估和解释, 去除多余的或不重要的模式, 将结果提交给用户。

KDD表示了从低层数据抽象到高级知识的过程。KDD过程必然是重复的, 数据挖掘的结果可能会要求在数据准备阶段作某些必要的变化, 模式的后处理也可能导致用户对模式类型作适当的修改等等。

二、常用安全审计方法

利用日志进行安全审计分析的思想, 最早是 1980年 Anderson的论文中正式提出的, 至今经历了 20余年的研究和发展, 已形成了较为完备的理论和应用体系。当前常用的安全审计方法主要有: 基于规则库的安全审计方法、基于数理统计的安全审计方法和基于数据挖掘的安全审计方法。

1. 基于规则库的安全审计。基于规则库的安全审计方法是将已知的攻击行为进行特征提取, 把这些特征用脚本语言等方法进行描述后放入规则库中, 当进行安全审计时, 将

收稿日期〕2005-04-20

作者简介〕石 彪 (1976-), 男, 湖南花垣人, 湖南商学院计算机与电子工程系教师, 硕士; 胡华平 (1967-), 男, 江西临川人, 国防科技大学计算机学院教授, 博士后。

收集到的审核数据与这些规则进行某种比较和匹配操作（关键字、正则表达式、模糊近似度等），从而发现可能的网络攻击行为。这种方法和某些防火墙和防病毒软件的技术思路类似，检测的准确率都相当高。基于规则库的安全审计方法有其自身的局限性。对某些特征十分明显的网络攻击行为，该技术的效果非常好；但对于其他一些非常容易产生变种的网络攻击行为（如 Backdoors等），规则库就很难满足要求。

2. 基于数理统计的安全审计。数理统计方法就是首先给对象创建一个统计量的描述，比如一个网络流量的平均值、方差等等，统计出正常情况下这些特征量的数值，然后用来对实际网络数据包的情况进行比较，当发现实际值远离正常数值时，就可以认为是潜在的攻击发生。数理统计方法的最大问题在于如何设定统计量的“阀值”，也就是正常数值和非正常数值的分界点，这往往取决于管理员的经验，不可避免产生误报和漏报。

3. 基于数据挖掘的安全审计。基于规则库和数理统计的安全审计方法已经得到了广泛应用，且获得了较大成功，但是它最大的缺陷在于已知的入侵模式必须被手工编码，它不能适用于任何未知的入侵模式。因此最近人们开始越来越关注带有学习能力的数据挖掘方法，目前该方法已在一些网络入侵检测系统得到了应用，它的主要思想是从“正常”的网络通信数据中发现“异”的网络通信模式，并和常规的一些攻击规则库进行关联分析，达到检测网络入侵行为的目的。将数据挖掘技术应用于安全审计已经成为一个研究热点，在这个领域已经有了近百篇论文。但是真正实现这样一套系统的还不多见。国外这方面做比较深入研究的主要有 Columbia University的 Wenke Lee研究组和 University of New Mexico（UNM）的 Stephanie Forrest研究组。国内这方面的研究则刚刚起步，中国科学院的国家信息安全重点实验室、东北大学国家软件工程研究中心、国防科技大学计算机学院等走在前列。

国内外大量的实验和测试结果表明，将数据挖掘技术应用于安全审计在理论上是可行的，在技术上建立这样一套系统是可能的。其技术难点主要在于如何根据具体应用的要求，从我们关于安全的先验知识出发，提取出可以有效地反映系统特性的特殊属性，应用合适的算法进行挖掘。技术难点还在于结果的可视化以及如何将挖掘结果自动地应用到实际的入侵检测系统中。

三、应用数据挖掘技术构建的安全审计引擎

计算机日志记录了计算机系统发生的各种重要事件，通过它可以了解系统运行状况，审核安全事件，诊断差错异常等。对日志的监控、审核与分析是系统管理员一项非常重要的工作。然而随着网络规模的扩大，网络设备数量的增多，日志的监控和审计范围也由原来单一的主机系统扩大到由各种服务器、路由交换设备和安全设备等组成的整个网络系统。网络系统中日志种类繁多、格式差别巨大，彼此之间缺少关联性，大小呈级数增长。海量日志数据的转存、归档、备份和分析将给系统管理员带来前所未有的工作压力。同时激增的日志数据背后隐藏着许多重要的信息，往往需要对其进行更高抽象层次的分析，以便更好地利用这些数据。

为了实现网络日志数据的综合分析和智能安全审计，本文采用了具有学习能力的数据挖掘技术来构建数据分析与安全审计引擎。其主要思路是首先收集足够多的“正常”或者“非正常”的历史日志数据，然后用一个分类算法去产生一个“判别器”来对新产生的待审计日志数据进行判别，决定哪些是正常行为而哪些是可疑或者入侵行为。此外系统还通过相关性分析找出被审计数据间的相互关联，通过时间序列分析算法来建立安全审计系统的时间顺序标准模型。图2描述了运用数据挖掘技术来实现的安全审计引擎框架。

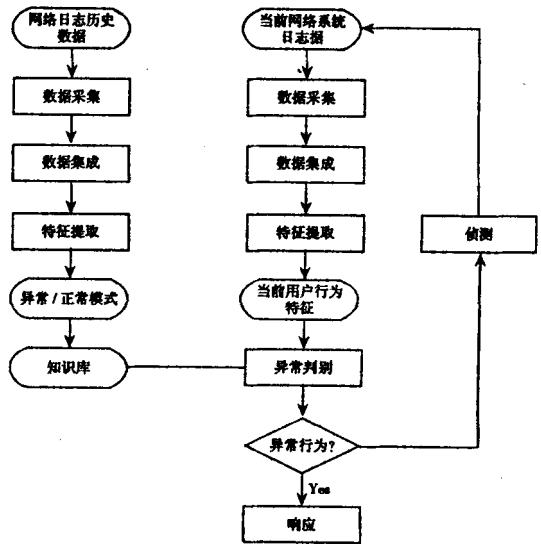

图 2　安全审计引擎框架

在图 2中，数据采集是收集网络中能够反映用户行为特征或能够描述网络行为状态的日志与审核数据，如主机系统日志、网络设备日志（交换机、路由器）、安全设备日志（防火墙、入侵检测系统）等；数据集成是将采集到的日志与审核数据进行集成与预处理，为下一步的数据挖掘准备数据，该部分主要是将多源的异构数据进行合并处理，解决语义的模糊性；特征提取是运用数据挖掘算法从预处理过的数据中，提取出能够反映网络正常或异常状态的特征，便于进一步抽象出构造知识库所需的正常或异常模式；知识库指知识库中存有安全审计引擎需要的网络行为正常/异常判定模式，安全审计引擎将新侦测到网络行为特征与其进行比较判断，从而可以判断出的网络行为是否是异常，进而判定是否存在攻击行为。该引擎的审计过程是：首先从知识库中提取出相关规则对反映网络行为的日志数据进行检测，根据检测结果做出相应处理。如属于异常行为，则引擎将通过邮件、铃声和系统短消息等方式通知管理员干预，并采取相应的缓解措施，如关闭系统服务，自动备份日志等。如属于正常行为，则将继续对新产生的网络日志进行监控和审计。本文中安全审计引擎系统原型的具体实现是以 Weka为基础。Weka是由新西兰 Waikato大学开发的并以 GNU方式发布的开源软件，它包含了一系列用于数据挖掘的机器学习算法，可以用来分析数据集，找到有用的模式。Weka采用 Java语言编写，提供了开放接口，可以用来开发自己的机器学习方案。

（下转第 101页）

盾的《林家铺子》、《子夜》；曹禺的《雷雨》、《日出》；老舍的《骆驼样子》等作品，将中国商业文学推向了前所未有的高峰。当代商业文学（中华人民共和国建国以来）主要集中在20世纪90年代以来这一特定的历史时期。中国商业经济前所未有的发展，与之相适应的文学创作也得到了蓬勃的发展；然而跟20世纪二、三十年代的商业文学相比较，中国当代商业文学的历史高度、现实深度、审美价值和艺术成就都大为逊色。

此外，与会者还就如何利用商业文学研究对高校（尤其是商科院校）课程建设和市场经济时代人才的培养的意义进行了探讨。有代表提出商业文学中有丰富的商业经验和良好的经营策略，商业文学能以其独特的艺术魅力滋养学生健康的商业精神，在高校开设商业文学课，有利于培养适应市场经济时代需要的新型人才；商科院校则可将商业文学以案例形式进入商学主干课程，利于商学开展案例教学；还可考虑独立设科、设系，将商业文学作为商学专业课，为培养专业商业作家、商务秘书、商业经纪人服务。

湖南商学院将“中国商业文学”列为重要的专题研究方向，在该院责任教授、学术带头人陈书良的带领下逐步形成了较为集中的科研群体，成立了全国首家“商业文学研究所”，出版了《中国商业文学发展概论》、《南宋江湖诗派与儒商思潮》、《现代商业社会的文学时尚》等专著，江苏社会科学院的萧相恺认为，这些成果填补了中国商业文学研究的空白，初步开拓了这一领域的研究。这次湖南商学院举办中国商业文学研讨会，汇集全国各地对此有兴趣、有研究的各位专家学者共同研讨中国商业文学的有关问题，有利于将中国商业文学研究引向深入，从而开阔中国文学的研究视野、拓展中国文学的研究领域。对此，与会代表都表示肯定与赞许。据悉，湖南商学院将牵头成立“中国商业文学学会”，以期扩大商业文学研究的影响，吸引更多同仁加入商业文学研究的行列。

（责任编辑：周小红）

（上接第85页）

四、结束语

本文在介绍数据挖掘技术和常用安全审计方法基础上，给出了一种基于数据挖掘技术实现的安全审计引擎框架。该安全审计引擎可从海量日志数据中提取安全特征，建立正常/异常模式知识库，然后根据知识库对新产生的日志数据进行异常判别，发现异常网络行为，并自动做出响应。

参考文献〕

[1] 田盛丰，黄厚宽. 人工智能与知识工程 [M]. 北京：中国铁道出版社，1999.

[2] Anderson J P. Computer Security Threat Monitoring and Surveillance. Fort Washington, James P. Anderson Co., 1980.

[3] 蒋嶷川，田盛丰. 入侵检测中对系统日志审计信息进行数据挖掘的研究 [J]. 计算机工程，2002，(1).

[4] T. Fawcett and F. Provost Combining data mining and machine learning for effective user profiling [A]. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining [C], Portland, OR, AAAI Press, 1996.

[5] http://www.cs.waikato.ac.nz/ml/weka/.

（责任编辑：周小红）