第32卷 第23期 计 算 机 工 程 2006年12月
Vol.32 No.23 Computer Engineering December 2006
·安全技术· 文章编号：1000—3428(2006)23—0159—03 文献标识码：A 中图分类号：TP311

WINEPI 挖掘算法在入侵检测中的应用

李永忠，孙 彦，罗军生
（江苏科技大学电子信息学院，镇江 212003）

摘 要：针对入侵检测系统的特点，将序列模式挖掘算法应用于入侵检测系统中。分析了 WINEPI 算法，并通过一个示例描述了数据挖掘的具体过程，讨论了挖掘结果的解释评估。
关键词：入侵检测；数据挖掘；序列模式

Application of WINEPI Mining Algorithm in IDS

LI Yongzhong, SUN Yan, LUO Junsheng
(College of Electrics and Information, Jiangsu University of Science and Technology, Zhenjiang 212003)

【Abstract】 According to the characteristics of IDS, the sequence pattern data mining algorithm is applied to IDS in this paper. The algorithm of WINEPI in IDS is presented, and the instance of using WINEPI algorithm is also presented. The results of sequence pattern data mining are explained.
【Key words】 Intrusion detection; Data mining; Sequence pattern

近年来，伴随着 Internet 规模的飞速发展，黑客攻击的频度、传播速度、受害面和穿透深度都在加大。网络安全问题已经成为影响网络发展、特别是商业应用的主要问题，并直接威胁着国家和社会的安全，受到了世界各国的广泛重视。如何建立安全而又健壮的网络系统，保证重要信息的安全性，已经成为计算机网络和通信领域研究的焦点。入侵检测技术是一种利用入侵留下的痕迹信息来有效地发现来自外部或内部的非法入侵的技术。入侵检测为系统提供了实时保护，被认为是防火墙之后的第 2 道安全闸门。

数据挖掘方法被Wenke.lee用于入侵检测中[1,2]，用数据挖掘程序处理搜集到的审计数据，为各种入侵行为和正常操作建立精确的行为模式，这是一个自动的过程，不需要人工分析和编码入侵模式。数据挖掘方法的关键点在于算法的选取和建立一个正确的体系结构。根据DARPA评估，运用数据挖掘方法的入侵检测系统在性能上优于基于“知识工程”的系统。

本文主要研究用 WINEPI 序列模式挖掘算法进行计算机网络入侵检测。

1 WINEPI 序列模式挖掘算法

1.1 序列模式挖掘

序列模式挖掘是在多个有序事件序列中查找出现频率大于某个阈值的序列模式的数据挖掘方法。序列模式挖掘方法侧重于数据之间的前后序列关系。它能发现数据库中形如“在某一段时间内，顾客购买商品 A，接着购买商品 B，而后购买商品 C，即序列 A→B→C 出现的频度较高”之类的知识，序列模式分析描述的问题是：在给定交易序列数据库中，每个序列是按照交易时间排列的一组交易集，挖掘序列函数作用在这个交易序列数据库上，返回该数据库中出现的高频序列。在进行序列模式分析时，需要由用户输入最小置信度 C 和最小支持度 S。

以往数据挖掘方法在网络安全方面应用的相关研究中，大部分都采用频繁模式挖掘算法。这种数据挖掘算法主要寻找给定数据集中项之间的有趣关系，可以用来发现诸如“如果一个顾客购买了计算机，那么他也倾向于购买计算机软件”这样的规则知识。然而这种数据挖掘方法没有考虑项之间序列关系和时间约束性。在网络入侵报警事件中，报警事件具有明显的时间序列性特征，这些事件在时间上有着先后顺序，逻辑上互为因果。正是基于入侵报警事件之间的这种相互序列关联的性质，所以序列模式挖掘方法更加适用于网络入侵事件之间的关联规则的发现。

入侵检测系统的关联性研究提出了一种通过入侵报警事件之间的关联性分析，通过手工或者自动的方法对入侵报警事件进行再组织和再分析，获得一个精简可用的报警集。在入侵检测系统的关联性研究中，一个关键的难点就是关联知识的获取，即发现可用于关联入侵报警的规则和知识。

1.2 WINEPI 算法

本文中序列模式挖掘采用串行WINEPI算法，这种算法是从电信网络故障管理中的报警关联系统中发展起来的[4]，其基本思想是首先找到短的频繁情景，然后逐步递推地找到大的频繁情景，在实际应用中表明该算法比较高效，适合于入侵检测关联规则发现，以下是对该算法的描述。

(1)算法描述

输入：入侵检测系统一段时间内产生的所有报警，指定时间窗口宽度 W，频繁度阈值 min_fr。

输出：入侵事件序列组成的集合，集合中每个元素具有不同的源地址目的地址对。

具体算法步骤如下：

基金项目：江苏省教育厅、江苏科技大学课题资助项目(2005DX006J，2005-2007)“计算机网络信息处理技术研究”
作者简介：李永忠(1961－)，男，硕士、教授，主研方向：计算机应用，网络安全，计算机信息处理；孙 彦、罗军生，硕士生
收稿日期：2006-02-17 **E-mail：**Liyongzhong61@163.com

Step1 把所有的入侵报警事件按照源地址目的地址对分类；

Step2 在每个分类上按照时间顺序升序排列；

Step3 在序列上滑动时间窗，统计每个时间窗上的所有情景；

Step4 计算每一个情景的频繁度

$$fr(a,s,win)=\frac{|w\in W(s,win)|a在W中|}{|W(s,win)|}$$

选出大于人工事先预定的阈值 min_fr 的所有情景，即为频繁情景，一个事件序列上的所有频繁情景组成一个频繁情景集 $F(s,win,\min\_fr)$；

Step5 找出所有的事件序列上的频繁情景集。

一旦发现了频繁情景集，就能很容易地产生序列模式，只需要分别统计所有的规则模式的置信度，然后选择输出置信度大于人工设定的置信度阈值的模式规则就可以了。

(2)WINEPI 在入侵检测中的具体应用

WINEPI 是一种有效的序列模式挖掘算法，但是作为一个通用的数据挖掘算法，网络入侵检测领域的特点并没有被考虑进去。在实际应用该算法的过程中，解决以下几个问题。

1)将时间滑动窗改变为事件滑动窗：由于在网络入侵检测中，网络入侵事件发生的相对时间先后关系比事件发生的绝对时间更加重要，而且数个事件之间的相对关系比一个时间段内所发生的事件更具关联性，因此改进 winepi 算法，将时间滑动窗改为事件滑动窗。

2)事件滑动窗宽度的选择：滑动窗宽度的选择是一个经验问题，宽度选择的太窄，挖掘出的特色情景就很少，而且算法的抗干扰能力下降，容易遗漏有特色的情景。宽度选择的太宽，挖掘出的情景太杂，容易将无关的事件关联在一起，而且降低了情景出现的频繁度，难以筛选。根据实际的经验，4~8 个事件是比较合适的滑动窗口长度。

3)偏序关系的选择：在经典的 WINEPI 算法当中，要求组成情景的事件之间存在着偏序关系，这种偏序关系是很宽泛的，可以是无序的，也可以是全序的。结合网络入侵事件检测领域的报警事件的特点，情景有可能是由一些事件按照严格的时间顺序组成的，例如网络入侵报警事件之间存在着时间上的先后顺序关系，逻辑上互为因果，因此选择按照事件发生时间的先后组成的全序关系是比较合理的，也有可能是由一些无序的事件组成的，例如一个扫描工具扫描不同的主机端口。

下面是用 DARPA2000 的报警数据，说明序列模式挖掘的具体过程。具体算法步骤如下：

Step1 把所有的入侵报警事件按源地址目的地址对分类。来自同一个 IP 的所有报警数据被分为一组，形成一个事务。

Step2 把所有的事件按照发生的时间升序排列，形成一个事件序列。本例中的序列总共包括 31 个报警事件，为便于说明，抽取其中的一段包含 10 个报警事件如表 1 所示，该序列基本上完整地表示了整个攻击的有效过程，这里只列出了报警类型和报警时间两个字段。为了便于描述，每一种报警类型提供一种符号表示。

Step3 用一个事件窗口长度为 4 的滑动窗口，在该序列上滑动形成一些列的窗口，如图 1 所示。

宽度为 w 的滑动窗口，在长度为 L 的时间序列上滑动，最终会形成(L+w-1)个窗口，例如，图 1 中总共形成了 13 个窗口。

下一步根据网络报警事件的性质统计每个窗口上包含的情景。这里的偏序关序选择了全序，统计结果如表 2 所示。

表 1 网络攻击报警事件表

报警类型	报警时间	符号
Sadmind Ping	10:48:19	A
TELNET access	10:48:22	B
TELNET login incorrect	10:48:26	C
Sadmind Ping	10:48:28	A
TELNET access	10:48:30	B
TELNET login incorrect	10:48:34	C
Sadmind Ping	10:58:19	A
TELNET access	10:58:20	B
RSERVICES rsh root	11:05:09	D
MStream_Zombie	11:06:18	E

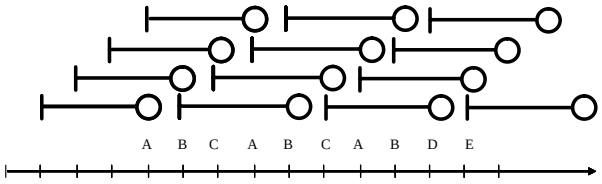

图 1 滑动窗口在事件序列上形成的子窗口

表 2 滑动窗口上形成的所有情景

窗口包含的序列	情景
A	{A}
A, B	{A}{B}{A,B}
A,B,C	{A}{B}{A,B}{A,C}{B,C}{A,B,C}
A,B,C,A	{A}{B}{C}{A,B}{A,C}{B,C}{B,A}{C,A}{A,B,C}{A,B,A} {A,C,A}{B,C,A}{A,B,C,A}
B,C,A,B	{A}{B}{C}{B,C}{B,A}{C,A}{C,B}{A,B}{B,C,A}{B,C,B} {B,A,B}{C,A,B}{A,B,C,A}
C,A,B,C	{A}{B}{C}{C,A}{C,B}{A,B}{A,C}{B,C}{C,A,B}{C,A,C} {C,B,C }{A,B,C}{ C,A,B,C}
A,B,C,A	{A}{B}{C}{A,B}{A,C}{B,C}{B,A}{C,A}{A,B,C}{A,B,A} {A,C,A}{B,C,A}{A,B,C,A}
B,C,A,B	{A}{B}{C}{B,C}{B,A}{C,A}{C,B}{A,B}{B,C,A}{B,C,B} {B,A,B}{C,A,B}{A,B,C,A}
C,A,B,D	{A}{B}{C}{D}{C,A}{C,B}{C,D}{A,B}{A,D}{B,D}{C,A,B} {C,A,D}{C,B,D }{A,B,D}{ C,A,B,D}
A,B,D,E	{A}{B}{D}{E}{A,B}{A,D}{A,E}{B,D}{B,E}{D,E}{A,B,D} {A,B,E}{A,D,E}{B,D,E}{A,B,D,E}
B,D,E,	{B}{D}{E}{B,D}{B,E}{D,E}{B,D,E}
D,E	{D}{E}{D,E}
E	{E}

在表 2 的基础上，统计每一个情景的频繁度，表 3 说明了其中几种感兴趣的情景模式的频繁度，其中频繁度等于包含该情景的窗口数除以所有的窗口数。

表 3 情景的频繁度

情景	频繁度
{B}	76.9%
{B,C}	46.2%
{A,B,C}	30.8%

Step4 计算输出情景规则的支持度，如表 4 所示。

在计算过程之中，还会输出其它一些序列模式，然而大多数都是没有意义的，所以必须经过进一步人工筛选，才能获得真正有意义的知识规则。根据表 3 和表 4 可以输出 2 个序列模式规则。

1)telnet access => telnet login error [4] (46.2%,60.1%)表示在 4 个事件之内，telnet access 导致 telnet login error 的可能性是 60.1%，该序列发生的频率是 46.2%。

2)Sadmind Ping => {telnet access, telnet login error} [4] (30.8%, 66.7%)表示在 4 个事件之内，Sadmind Ping 导致 telnet

尝试的可能性是 66.7%，该序列发生的频率是 30.8%。

表 4 输出规则的支持度

情景规则	支持度
{B}->{B,C}	60.1%
{A}->{A,B,C}	66.7%

经过分析，这个情景对应着 Sadmind Buffer Overflow 攻击，黑客首先连接 Sadmind 服务端口，试图执行缓冲溢出代码，在系统中添加一个新账号，触发了 Sadmind_Ping 报警，接着黑客尝试用该帐号登录系统，以确认溢出攻击是否成功，所以触发了 telnet access 和 telnet login error 报警。从报警序列中可以看出，该黑客经过两次失败的尝试之后，终于在第 3 次成功地攻克了该主机。

在实际应用中，事件滑动窗口的宽度、置信度阈值都会影响到数据挖掘的结果，因此在程序设计中，允许手动调整事件滑动窗口的宽度和置信度阈值。

(3)算法总结

在实际的应用中，序列模式挖掘出的特色情景有很多，其中对网络入侵事件分析比较有用的有以下几种。

1)和攻击工具有关的特色情景 Episode：一个源 IP 地址产生了一系列的针对不同目标的报警事件，这说明该黑客使用了自动化的攻击工具。

2)有一些序列模式队则的推理右件代表了一系列的攻击，而其推理左件正好预测了这一系列攻击的发生。

3)有一些报警事件 A 总是必然地导致另外一些报警事件 B 在一定时间之内接连发生，但是反过来并不成立。即在一定时间内，报警事件 A 总是百分之百导致报警事件 B 的发生，而报警事件 B 并不会导致报警事件 A 百分之百发生。

4)有一些合法的系统操作会引起报警，例如某些远程系统调用或者网络管理功能。

1.3 结果的解释评估

对上述知识发现过程中发现的关联规则和序列模式进行分析，发现有两点规律。

(1)频繁的模式并不一定有意义。有些频繁模式只是统计意义上的频繁模式，实际上并没有意义，所以这样的规则可以予以剔除。

(2)支持度和置信度低的关联规则和序列模式不一定没有意义。有些入侵事件发生的次数很少，而且规律性不强，产生的规则的支持度和置信度都很低，但这样的规律对于入侵事件关联很重要。

由此看出，如果仅按照情景频繁度和模式置信度对 KDD 发现的规则进行筛选，必然会保留很多频繁而没有意义的模式，同时一些有价值但不频繁的模式就没法保留下来。

为了进一步对以上数据挖掘的结果进行自动地解释评估，减少数据挖掘过程输出的模式规则的数量，就必须引进新的过滤指标，按照一定的阈值进行过滤，然后再输出规则知识。这样一来，不但可以减少模式规则的数量，同时这样的模式规则也具有更高的可信度和可理解性。

2 仿真试验

为了提高入侵检测的评估和预测能力，美国国防部(Defense Advanced Research Projects Agency)资助 MIT 的林肯实验室(Lincoln Laboratory)进行入侵检测的评估。为此，林肯实验室每年会拿出一部分实验室产生的攻击数据作为入侵检测的数据源。本次实验就是采用的其发布的 2000 年的数据(DARPA2000)，这一年的数据包括了一个完整的 DDOS 的攻击流程。

Snort 和 Real Secure 两种检测器总共上报了 1 102 条报警数据。在这样的数据集上，首先按照交互式知识发现的方法寻找频繁模式，利用人工经验控制知识发现的过程，评价知识，删除无意义的知识，加入有意义的知识。表 5 为知识发现算法、采用的参数和最终发现的模式个数以及交互式处理后的模式个数在这样的数据集上，所有这些知识都可以应用在入侵事件的分析之中。

表 5 算法参数及发现的知识统计

WINEPI 算法参数	滑动窗口长度＝6　频繁度门限＝10% 置信度阈值＝70%
频繁情景数	34
输出情景规则数	5

交互式手工处理后的模式可以转化成入侵事件关联规则加入专家系统的规则库中，将这些规则和以前的规则库中已有的规则用于这段入侵报警数据集的关联，最终得到的数据集，报警事件数目约简到 451 条。

3 总结

本文通过一个示例详细描述了 WINEPI 数据挖掘的具体过程，并讨论了数据一致化处理方法。经过仿真试验验证了算法的有效性。实验结果表明本文所述的关联分析方法确实可以有效地降低报警事件的数量。

参考文献

1 Weeke L. A Data Mining for Constructing Feature and Model for Intrusion Detection System[D]. Columbia: Columbia University, 1999.

2 Weeke L, Stolfo J, Mok K W. Algorithms for Mining System Audit Data[C]. Proceedings of the IEEE Symposium on Security and Privact, 1999.

3 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

4 Peng N, Yun C, Reeves D S. Analyzing Intensive Intrusion Alerts via Correlation[C]. Proc. of the 5th International Symposium on Recent Advance in Intrusion Detection, Zurich, Switzerland, 2002.

（上接第 147 页）

频水印研究能达到预期目的。本方法适用于音频字信号的版权保护及信息隐藏等领域。

参考文献

1 Xin L, Heather Y H. Transparent and Robust Audio Data Hiding in Cepstrum Domain[C]. Proceedings of IEEE International Conference on Multimedia and Expo 2000, New York, 2000: 397-400.

2 Kwang L S, Sung H Y. Digital Audio Watermarking in the Cepstrum Domain[J]. IEEE Trans. on Consumer Electronics 2000, 46(3): 744-750.

3 张敏瑞, 易克初. 倒谱域音频与图像水印算法[J]. 西安电子科技大学学报, 2003, 30(6): 730-738.