基于警报序列聚类的多步攻击模式发现研究

梅海彬[1,2]，龚俭[1]，张明华[2]

（1. 东南大学 计算机科学与工程学院 江苏省计算机网络技术重点实验室，江苏 南京 210096；
2. 上海海洋大学 信息学院，上海 201306）

摘 要：研究了从警报数据中发现多步攻击模式的方法。通过定义警报间的相似度函数来构建攻击活动序列集。采用序列比对技术，将具有相似攻击行为的序列进行聚类。基于动态规划的思想，通过抽取最长公共子序列的算法自动发现类中的多步攻击模式。该方法不需要依赖大量先验知识，设置参数少，易于实现。实验结果验证了该方法的有效性。

关键词：入侵检测；警报关联；多步攻击；聚类

中图分类号：TP393.08 **文献标识码**：A **文章编号**：1000-436X(2011)05-0063-07

Research on discovering multi-step attack patterns based on clustering IDS alert sequences

MEI Hai-bin[1,2], GONG Jian[1], ZHANG Ming-hua[2]

(1. Computer Network Technology Key Laboratory of Jiangsu Province, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China; 2. College of Information, Shanghai Ocean University, Shanghai 201306, China)

Abstract: A method of discovering multi-step attack patterns from alert data was studied. Alert similarity function was defined to construct the set of attack activity sequences. Sequence alignment technology was used to cluster the similar attack activity sequences. Multi-step attack patterns in a cluster were automatically discovered by the longest common subsequence extraction algorithm based on the idea of dynamic programming. The proposed method didn't depend on large amounts of prior knowledge. Few configuration parameters were needed and it was easy to implement. Experimental results demonstrate the effectiveness of proposed method.

Key words: intrusion detection; alert correlation; multi-step attack; clustering

1 引言

根据国家计算机网络应急技术处理协调中心[1]的年度网络安全工作报告，近几年来大部分攻击尤其是危害大的攻击几乎都是复杂的多步攻击。发现攻击者的多步攻击模式，可以深入了解攻击者的攻击动机、经常使用的攻击技巧等，从而可帮助安全管理者构建一个更加完善的安全防护体系。此外，多步攻击模式发现在攻击意图推测、攻击响应、警报关联以及网络攻击取证等领域也具有重要的意义。然而，目前大部分入侵检测系统（IDS, intrusion detection system）的检测引擎实现简单，无法描述

收稿日期：2010-08-06；**修回日期**：2010-12-09

基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2009CB320505）；上海高校选拔培养优秀青年教师科研专项基金资助项目（ssc09015）

Foundation Items: The National Basic Research Program of China (973 Program) (2009CB320505); The Scientific Special Funds for Cultivation and Selection of Excellent Young High Education Teachers of Shanghai (ssc09015)

相对复杂的多步攻击特征，且缺乏必要的关联能力，这使得IDS只能对单独的攻击行为报警，存在警报数据量大且不易理解和分析，无法识别多步攻击行为的不足。因此，非常有必要研究如何从大量的警报数据中提取更深层次的多步攻击模式，以掌握入侵的全貌。

由多步攻击的过程可知[2]，攻击者为了达到特定的攻击目的，不同攻击步骤间的发生次序具有一定的模式，这导致采用相同攻击策略的攻击者，在攻击行为上具有相似性。此外，网络中存在许多自动攻击工具软件或能自动传播的恶意软件，它们的攻击模式基本固定，在攻击行为上也具有相似性。这些攻击行为的相似性会在IDS产生的警报中表现出来，所以通过对警报的分析可以发现相似攻击行为中的多步攻击模式。现有的警报分析方法虽然能够识别多步攻击模式，但大多需要使用大量先验知识定义复杂的规则，或需要配置过多难以确定的参数，实用性不高。据此，本文提出了一种基于警报序列聚类自动从警报数据中发现多步攻击模式的新方法。方法不需要预先指定复杂的先验知识，难以配置的参数少，从而易于实现。实验结果表明，本文方法能有效识别警报数据中的多步攻击模式。

本文第2节介绍了国内外相关研究；第3节详细描述基于警报序列聚类的多步攻击模式发现方法；第4节通过实验验证方法的有效性；最后是结束语。

2 相关工作

目前，研究者已经提出了多种通过警报关联分析技术发现多步攻击模式的方法。Peng Ning 等[3]利用攻击之间的依赖关系提出了具有代表性的基于前因后果的警报关联方法。该方法的关键在于如何准确定义每种攻击的前提与结果，这需要大量的专家先验知识。此外，由于方法完全依赖手工定义的关联规则，因此不能发现包含有新的攻击或攻击关系的攻击模式。

针对以上不足，Xinzhou Qin 等[4, 5]提出了基于统计的警报关联方法。方法引入时间序列模型来计算警报时间序列变量之间的 Granger 因果指数（GCI，granger causality index），如果GCI的值大于给定的阈值，则判定两警报具有Granger因果联系，可进行关联。方法不需要过多的先验知识，可以发现新的多步攻击模式。但需要为每对警报进行时间序列转换和判定是否具有 Granger 因果联系，计算量大，需要配置的参数较多[6]。Bin Zhu 等[7]还提出了基于多层感知器和支持向量机的警报关联方法，但方法需要使用训练数据，而在安全领域准备符合要求且足够多的训练数据往往很困难。

在此基础上，Zhitang Li 等[8]提出一种新的基于数据挖掘发现多步攻击模式的方法。该方法虽然不需要过多的先验知识，但是采用固定的时间窗口值来划分警报子序列。由于每种多步攻击模式的持续时间不同，因此，时间窗口值往往较难确定。此外，方法在划分警报序列时仅使用了警报产生的时间和攻击类型2个属性，未考虑警报IP地址这一重要信息，影响了挖掘的正确性。

文献[9]和文献[10]同样利用数据挖掘的方法来发现多步攻击模式。文献[9]采用了一个用户自定义的固定时间窗口对警报进行聚集，以形成挖掘频繁图模式的候选项。针对固定时间窗口值不够灵活的问题，文献[10]提出了扩展的事件时间窗口，但引入了2个新的参数，即时间窗口间的间隔和最大扩展事件时间窗口值。这些参数的设置非常重要，但主要依赖于多次实验和专家的经验，因此不易确定。

可见，采用数据挖掘的多步攻击模式发现方法能够减少对先验知识、训练数据的依赖，具有发现新型攻击模式的能力，但现有的方法在划分警报子序列，攻击模式挖掘的准确性等方面还存在不足，影响了方法的实用性。

3 基于警报序列聚类的多步攻击模式发现

3.1 基本思路

方法首先将IDS警报数据转化为警报序列，并根据警报之间的相似度，构建攻击活动序列集。然后，采用序列比对技术来度量攻击活动序列间的距离，将具有相似攻击行为模式的序列聚集在一起。最后，基于动态规划的思想，通过求序列集的最长公共子序列来发现警报数据中的多步攻击模式。由于方法通过相似度来衡量警报属于同一多步攻击活动的可能性大小，从而弥补了文献[8]中简单地使用固定时间窗口对警报数据进行划分的不足，使得划分出的攻击活动序列更符合攻击活动发生的实际情况，构建的攻击活动序列集更准确。此外，方法在构建活动序列集时比文献[8]多考虑了警报的IP地

址信息，提高了发现多步攻击模式的准确性。最后，本文方法较文献[4, 5]设置的参数少，易于实现。

3.2　构建攻击活动序列集

多步攻击具有序列性、时间性等特征[11]，即一种特定的多步攻击的攻击步骤之间具有一定的发生模式，并且这些攻击步骤会在某个时间间隔内完成。为了从警报数据中发现多步攻击模式，本文首先将警报数据转换为可能包含这些多步攻击步骤的警报序列集合。

定义 1　警报序列。设 IDS 产生的警报 a 可用 m 元组$(p_1,p_2,\cdots,p_m)$表示，$p_k(1\leqslant k\leqslant m)$为 a 的第 k 个属性，并记 $AD=\{a_1,a_2,\cdots,a_{|AD|}\}$为警报集合，其中，$a_i(1\leqslant i\leqslant |AD|)$表示第 i 个警报。将 AD 中的警报按其产生的时间进行排列，所构成的序列称为警报序列，记为 $AS=<a_1,a_2,\cdots,a_{|AS|}>$，满足 $a_i.time\leqslant a_j.time$ $(1\leqslant i\leqslant j\leqslant |AS|)$。

定义 2　攻击活动序列。攻击活动序列是指由一次多步攻击活动所引发的警报序列，记为 $AT=<a_1,a_2,\cdots,a_n>$，其中，$a_i\ (1\leqslant i\leqslant n)$表示第 i 个警报，满足 $a_i.time\leqslant a_j.time\ (1\leqslant i<j\leqslant n)$。

多步攻击活动所对应的警报在时空属性上存在内在的联系。本文在文献[12]的基础上，定义了警报间在时间、IP 地址上的相似度以及总的相似度计算函数，分别为式(1)、式(2)和式(3)。当警报间具有较大相似性时，认为它们属于同一次攻击活动。

$$f_t(a,b)=\begin{cases}1, & \Delta t\leqslant t_{\min}\\ (t_{\max}-\Delta t)/(t_{\max}-t_{\min}), & t_{\min}<\Delta t<t_{\max}\\ 0, & \Delta t\geqslant t_{\max}\end{cases}\quad(1)$$

$$f_{ip}(a,b)=\begin{cases}1, & a.dst=b.src \text{ 且 } a.src\neq b.dst\\ \max\{r(a.type,b.type),sim(a.src,b.src)\}, & a.dst=b.dst\end{cases}\quad(2)$$

$$sim(ip_1,ip_2)=\frac{\max\{n\mid h_{\text{sub}}(n,ip_1)=h_{\text{sub}}(n,ip_2)\}}{32}$$

$$sim(a,b)=f_t(a,b)+f_{ip}(a,b),\ a.time\leqslant b.time\quad(3)$$

其中，a 和 b 为任意给定的警报，Δt 为$|a.time-b.time|$，$t_{\min}$ 和 $t_{\max}$ 是 2 个给定的阈值，$h_{\text{sub}}(i,ip)$表示取 ip 地址的高 i 位。$r(a.type,b.type)$为 2 种警报类型之间的关系值。

依据警报间总的相似度可以从警报数据中构建出攻击活动序列集，算法描述如下。

算法 1

input: raw alert dataset $AD=\{a_1,a_2,\cdots,a_{|AD|}\}$ and the threshold α;

output: alert sequences set ATS;

$ATS=\varnothing$; transform AD into alert sequence AS using alert timestamp;

for (each alert a_i in AS) {

$S_{tmp1}=0$; $S_{tmp2}=0$; $k=0$;

for ($j=0;j\leqslant |ATS|;j++$){

$S_{tmp2}=\max\{sim(a_i,b_{\text{m}})|\ b_{\text{m}}\in AT_{\text{j}}\}$ /*计算 a_i 与 AT_j 的相似度*/

if ($S_{tmp1}\leqslant S_{tmp2}$){ $k=j$; $S_{tmp1}=S_{tmp2}$; }

}

if ($S_{tmp1}\geqslant\alpha$) { add a_i to the end of AT_k; } else {

create a new attack seqence AT_{new};

add a_i to AT_{new};

$ATS=ATS\cup\{AT_{\text{new}}\}$;

}

}

return ATS

3.3　攻击活动序列聚类

根据聚类原理[13]，聚类时需要给定元素之间的距离测度或相似测度。对于序列元素，简单的方法是使用 Hamming 距离，但方法不够灵活，要求两序列长度相等且没有考虑两序列中元素之间位置的对应关系。考虑到攻击活动序列的复杂性，本文使用基于序列比对的距离计算方法[14]。

定义 3　序列操作类型集合。序列操作类型集合是指将一序列转换为另一序列可供选择的所有操作种类的集合，记为 Ξ。将一序列转换为另一序列所进行的一系列操作构成的序列称为操作序列，记为 $O=<o_1,o_2,\cdots,o_n>$，$o_i\in\Xi$。本文定义 $\Xi=\{ins(e,l),del(e,l),rep(e,e',l)\}$，其中，$ins(e,l)$表示插入操作，即在序列的第 l 位置插入 e 类型警报；$del(e,l)$表示在序列的第 l 位置删除 e 类型警报；$rep(e,e',l)$表示在序列中将位置为 l 的 e 类型警报替换为 e' 类型的警报。

定义 4　攻击活动序列距离。对于给定的攻击活动序列 AT_i 和 AT_j，使这 2 个序列成为相等序列所花费的最少操作代价称为这 2 个攻击活动序列的距离，记为 $dist(AT_i,AT_j)$，其值可由式(4)计算。

$$dist(AT_i,AT_j)=\min\{C(O_l)\mid AT_i\xrightarrow{O_l}AT_j\}\quad(4)$$

其中，$AT_i\xrightarrow{O_l}AT_j$ 表示使攻击活动序列 AT_i 与 AT_j 相等所需要的操作序列为 $O_l=<o_1,o_2,\cdots,o_k>$，每个操

作 $o_i(1\leqslant i\leqslant k)$的代价为 $c(o_i)$。$C(O_l)$为操作序列 O_l 的总操作代价，即 $C(O_l)=\sum_{i=1}^{k}c(o_i)$。

1) 距离计算

采用动态规划的思想[15]可计算序列之间的距离。设 $AT_1[i]$和 $AT_2[j]$分别表示序列 AT_1 和 AT_2 的第 i 个和第 j 个元素$(1\leqslant i\leqslant|AT_1|,1\leqslant j\leqslant|AT_2|)$，$s(i,j)$表示将序列 AT_1 的前 i 元素转变为序列 AT_2 的前 j 个元素所需要的最小代价，则有式(5)：

$$s(i,j)=\begin{cases}0, & i=0\ \text{且}\ j=0\\ s(i-1,j)+w_{\text{del}}, & i\geqslant 1\ \text{且}\ j=0\\ s(i,j-1)+w_{\text{ins}}, & i=0\ \text{且}\ j\geqslant 1\\ f(i,j), & i\geqslant 1\ \text{且}\ j\geqslant 1\end{cases}\tag{5}$$

其中，w_{del} 和 w_{ins} 分别代表删除和插入操作的代价，$k(i,j)$和 $f(i,j)$函数定义为式(6)：

$$\begin{aligned}&k(i,j)=\begin{cases}0, & AT_1[i]=AT_2[j]\\ w_{\text{del}}+w_{\text{ins}}, & AT_1[i]\neq AT_2[j]\end{cases}\\ &f(i,j)=\min\{s(i-1,j)+w_{\text{del}},\\ &s(i,j-1)+w_{\text{ins}},\ s(i-1,j-1)+k(i,j)\}\end{aligned}\tag{6}$$

根据式(5)，则攻击活动序列 AT_1 和 AT_2 的距离大小为 $dist(AT_1,AT_2)=s(|AT_1|,|AT_2|)$。

2) 聚类算法

为了尽量减少聚类的时间开销和相关的参数设定，本文采用一种简单且易于实现的聚类算法。算法描述如下。

算法 2

input: *ATS*, σ;

output: *cset*, the set of clusters after clustering;

cset = Ø;

for (each $AT_i\in ATS$) {

find cluster c_i, which satisfied $c_i=\arg\min_{c_j} f_{dist}(AT_i,c_j)$, $c_j\in cset$;

if ($f_{dist}(AT_i,c_i)$ <=σ) { $c_i=c_i\cup\{AT_i\}$; } else {

create a new cluster c_{new};

$c_{\text{new}}=c_{\text{new}}\cup\{AT_i\}$;

$cset=cset\cup c_{\text{new}}$;

}

}

return *cset*

其中，$f_{dist}(AT_i,c_i)$函数的返回值是序列 AT_i 与类 c_i 中各序列距离的最小值。算法基本思路与文献[16]中对警报进行融合的思路类似，均采用了贪婪算法思想，但文献[16]是将警报作为处理对象，距离定义为警报之间的相似度，目的是如何将新来的警报加入到已有的攻击场景中，而本文算法是对警报序列进行处理，距离是警报序列之间的相似度，其目的是将新来的警报序列聚集到已有聚好的类中。所以算法与文献[16]在处理对象以及目的上存在不同，本算法是将多个具有内在联系的警报作为一个整体进行处理，其意义在于能将警报数据中具有相似攻击行为的警报序列进行聚类，以便自动发现这些警报序列中所蕴含的多步攻击的行为模式。

3.4 多步攻击模式的发现

本文采用求序列集的最长公共子序列的方法从聚类中发现多步攻击模式，以下为最长公共子序列的相关概念以及求最长公共子序列的算法。

定义 5 子序列。对于给定的序列 $S_1=<a_1,a_2,\cdots,a_m>$和 $S_2=<b_1,b_2,\cdots,b_n>$，如果 S_2 存在下标 $i_1<i_2<\cdots<i_m$，使得 $a_j=b_{i_j}$ $(1\leqslant j\leqslant m)$，则称 S_1 是 S_2 的子序列，记为 $S_1\prec_s S_2$。

定义 6 最长公共子序列。如果序列 S 同时满足 $S\prec_s S_1$ 和 $S\prec_s S_2$，则称 S 为 S_1 和 S_2 的公共子序列，记为 $S\prec_c(S_1,S_2)$。如果 $S'\prec_c(S_1,S_2)$，且不存在其他的公共子序列的长度大于 S' 的长度，则称 S' 为 S_1 和 S_2 的最长公共子序列(LCS, longest common subsequence)。

求序列 X 和 Y 的 LCS 的最直接方法是穷举法，但其时间复杂度为 $O(2^{\min(|X|,|Y|)})$，考虑到 LCS 问题具有最优子结构性质[17]，利用该性质可用动态规划的思想设计出求 LCS 的有效算法，使算法复杂度降为 $O(|X||Y|)$。算法 3 是求类 c 中 LCS 的算法描述，其中，函数 $f_{\text{LCS}}(X,Y)$表示求 X 和 Y 的 LCS。

算法 3

input: an attack activity sequence cluster *c*;

output: the LCS of attack activity sequence cluster *c*;

TC= *c*;

while (|*TC*|>1){ /*类 *TC* 中的元素个数大于 1*/

$\forall X,Y\in TC$ let $S=f_{\text{LCS}}(X,Y)$;

remove *X* and *Y* from *TC*;

insert *S* into *TC*;

}

return *TC*

4　实验与结果分析

4.1　基于 DARPA 2000 数据集的实验

DARPA 2000 数据集[18]是由 MIT 林肯实验室创建的目前最权威的攻击场景测试数据集，具有较强的代表性。数据集包含 2 个多步攻击的场景，分别为 LLDOS1.0 和 LLDOS2.0.2。实验时，本文使用 Peng Ning 等[19]利用 RealSecure 入侵检测系统在 LLDOS1.0 上产生的警报数据，共 1 813 条。表 1 为警报数据经冗余消除后数目位于前 5 的警报类型分布情况。

表 1　LLDOS1.0 中警报数目前 5 的警报类型分布

序号	警报类型	数量	比例/%
1	Email_Ehlo	724	57.41
2	TelnetTerminaltype	161	12.77
3	Email_Almail_Overflow	78	6.19
4	FTP_Pass	73	5.79
5	FTP_User	73	5.79

在此警报数据的基础上共得到 944 个攻击活动序列。这些序列中包含大量长度为 1 的序列（这说明警报中确实存在大量零碎警报[20]），删除长度为 1 的序列后共得到 195 个攻击活动序列。对这些攻击活动序列进行聚类和提取多步攻击模式后共得到 6 个序列模式，如表 2 所示。

表 2　　LLDOS1.0 中发现的序列模式

序号	序列模式
1	<FTP_User,FTP_Pass,FTP_Syst>
2	<TelnetTerminaltype,TelnetEnvAll,TelnetXdisplay>
3	<Email_Ehlo,TelnetTerminaltype>
4	<Email_Ehlo,Email_Debug>
5	< Email_Ehlo,TelnetTerminaltype>
6	<Sadmind_Ping,Admind,Sadmind_Amslverify_Overflow, Admind,Rsh,Mstream_Zombie>

根据 RealSecure 的规则说明文档以及测试数据集的描述文档，发现 1 到 5 号序列均属于正常行为模式，而 6 号序列对应了 LLDOS1.0 中真正的攻击步骤，但该序列多了 2 个 Admind 警报，少了对应于最后攻击步的 Stream_DoS 警报。产生原因是 Admind 警报太普遍，只要有事件涉及到 Sadmind 服务时 RealSecure 均会产生该警报[20]，实际上可以通过配置 RealSecure 规则将这个警报过滤掉。而缺少 Stream_DoS 警报的原因是该警报的源地址是伪造的 IP 地址。如果将序列中 2 个多余的 Admind 警报删除，则本文发现的多步攻击模式如图 1 所示。图 1 中的多步攻击模式采用一种被研究者广泛使用的有向无环图来表示，其中的结点表示多个攻击步骤中的一步，结点中的文字表示警报的类型，而连接 2 个结点的有向边表示了 2 个攻击步先后发生的依存关系。

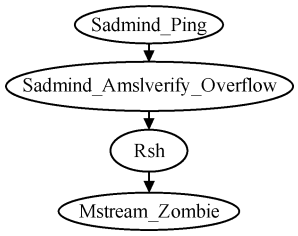


图 1　LLDOS1.0 中发现的多步攻击模式

4.2　基于 Monster 数据集的实验

为验证方法在真实 IDS 警报数据上的可行性，本文采用了 Monster 3.0 系统①在 CERNET 江苏省网主干上 1 个月的警报记录作为多步攻击模式挖掘的数据集。其中共有 1 016 265 条警报，208 种警报类型。表 3 为数目位于前 5 的警报类型分布情况。

表 3　　Monster 警报数目前 5 的警报类型分布

序号	警报类型	数量	比例/%
1	WEB-MISC /cgi-bin/// access	426 152	41.93
2	ATTACK-RESPONSES 403 Forbidden	265 057	26.08
3	ICMP Destination Unreachable (Communication with Destination Host is Administratively Prohibited)	60 838	5.99
4	WEB-MISC whisker HEAD with large datagram	36 591	3.60
5	ICMP PING speedera	34 553	3.40

从这些警报数据中共得到 76 440 条攻击活动序列，删除长度为 1 的序列后，共得到 11 941 条攻击活动序列。通过对这些活动序列的聚类和提取 LCS，共得到 142 个序列模式，限于篇幅，这里只列出出现频率较高的 5 个序列模式，见表 4。

表 4　实验中发现的出现频率较高的前 5 个攻击序列模式

序号	序列模式
1	<SCAN nmap fingerprint attempt,SCAN nmap XMAS,SCAN FIN,SCAN XMAS>
2	<WEB-MISC adminlogin access,WEB-IIS cmd.exe access, WEB-MISC whisker HEAD with large datagram, WEB-MISC backup access,WEB-MISC server-info access>
3	<INFO TELNET Bad Login,INFO FTP Bad login>
4	<WEB-IIS cmd.exe access,WEB-IIS unicode directory traversal attempt>
5	<ICMP L3retriever Ping,WEB-FRONTPAGE shtml.dll access>

① CERNET 华东北地区网络中心开发的面向大规模网的 IDS。

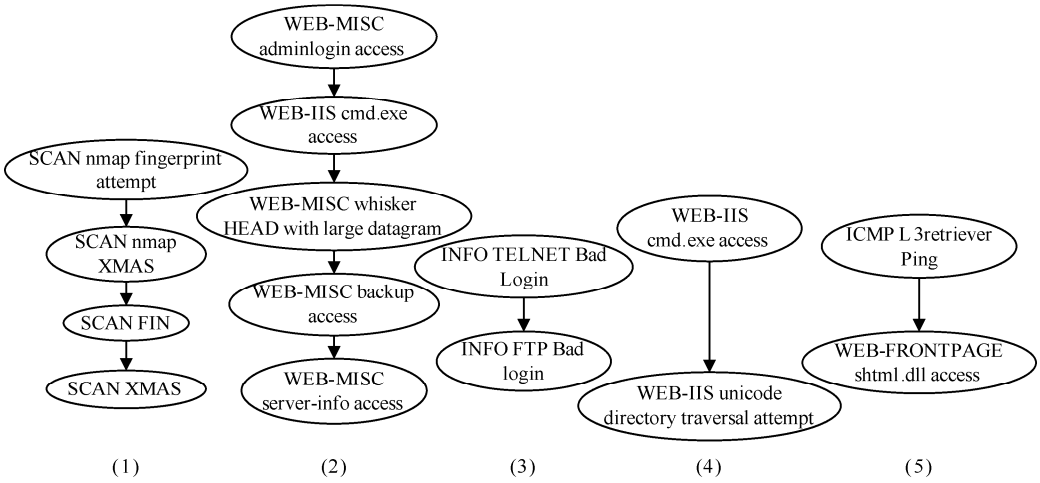
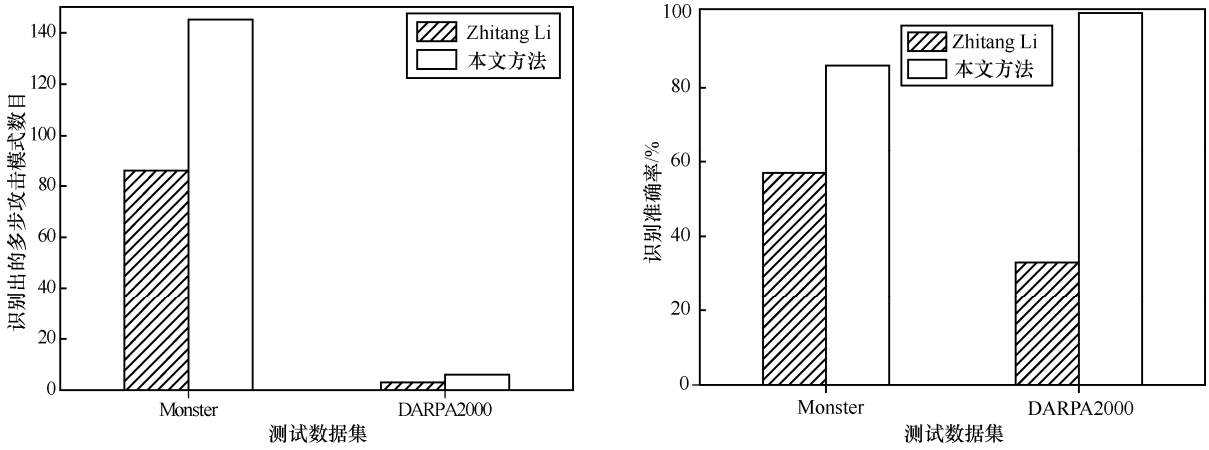


图 2 Monster 数据集中发现的多步攻击模式



(a) 识别的多步攻击模式数目 (b) 识别准确率

图 3 本文方法与 Zhitang Li 方法的实验对比

序列 1 显示了攻击者利用多种扫描方法进行端口扫描。序列 2 则是以多种方式对 Web 服务器进行入侵的尝试。同样，其他序列所对应的攻击基本也都是一种试探性攻击。这些序列所对应的攻击模式可以用图 2 表示。

从上述实验结果可以看出，本文所提的方法能够较好地识别警报数据中的多步攻击模式。由于通过聚类来挖掘相似攻击行为发生的公共序列模式，因此本文方法不需要事先给定复杂的关联规则和设定过多难以确定的参数，具有易于实现和能发现新的多步攻击模式的优点。

最后，考虑到本文方法与 Zhitang Li 等人的方法都是基于数据挖掘技术来发现多步攻击模式，为进一步验证本文方法的有效性，与他们的方法进行了实验对比。实验仍采用上述的 2 个测试数据集，并按文献[8]中的参数进行设置，实验结果如图 3 所示。从图中可以看出，本文方法在识别的多步攻击模式数目以及识别准确率两方面都具有优势。

5 结束语

本文依据多步攻击行为具有的序列特征，以及各攻击步骤之间具有特定模式的特点，提出了一种利用警报序列聚类自动从 IDS 大量琐碎警报中挖掘攻击模式的新方法。方法定义了警报之间相似度的计算函数，并给出了攻击活动序列划分、序列聚类以及模式挖掘的相应算法。由于方法不需要依赖预先手工生成的各种各样复杂的关联规则，也不需要配置过多难以确定的参数，与目前典型多步攻击模式发现方法相比，方法更易于实现，实用性更强。通过在经典测试数据集和实际大规模网 IDS 警报数据集上进行的相关实验，以及与其他方法的实验比较，验证了方法的有效性。

参考文献：

[1] 中国国家计算机网应急技术处理协调中心[EB/OL]. http://www.cert.org.cn/, 2010.
The national computer emergency response teams/coordination center of China[EB/OL]. http://www.cert.org.cn/, 2010.

[2] 鲍旭华, 戴英侠, 冯萍慧等. 基于入侵意图的复合攻击检测和预测算法[J]. 软件学报, 2005, 16(12):2132-2138.
BAO X H, DAI Y X, FENG P H, *et al*. A detection and forecast algorithm for multi-step attack based on intrusion intention[J]. Journal of Software, 2005,16(12):2132-2138.

[3] NING P, XU D. Learning attack strategies from intrusion alerts[A]. Proceedings of the 10th ACM Conference on Computer and Communications Security[C]. Washington, D C, USA, 2003. 200-209.

[4] QIN X, LEE W. Statistical causality analysis of INFOSEC alert data[A]. Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection[C]. Pittsburgh, USA, 2003. 73-94.

[5] QIN X, LEE W. Discovering novel attack strategies from INFOSEC alerts[A]. Proceedings of the 9th European Symposium on Research in Computer Security[C]. Sophia Antipolis, France, 2004. 439-456.

[6] MAGGI F, ZANERO S. On the use of different statistical tests for alert correlation: short paper[A]. Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection[C]. Gold Goast, Australia, 2007. 167-177.

[7] ZHU B, GHORBANI A A. Alert correlation for extracting attack strategies[J]. International Journal of Network Security, 2006, 3(3):244-258.

[8] ZHANG A, LI Z, LI D, *et al*. Discovering novel multistage attack patterns in alert streams[A]. Proceedings of International Conference on Networking, Architecture, and Storage[C]. Guilin, China, 2007. 115-121.

[9] SADODDIN R, GHORBANI A A. An incremental frequent structure mining framework for real-time alert correlation[J]. Computers & Security, 2009, 28(3, 4):153-173.

[10] WANG L, GHORBANI A, LI Y. Automatic multi-step attack pattern discovering[J]. International Journal of Network Security, 2010, 10(2):142-152.

[11] 王莉. 网络多步攻击识别方法研究[D]. 武汉: 华中科技大学, 2007.
WANG L. Study on Method of Network Multi-stage Attack Plan Recognition[D]. Wuhan: Huazhong University of Science and Technology, 2007.

[12] VALDES A, SKINNER K. Probabilistic alert correlation[A]. Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection[C]. Davis, CA, USA, 2001. 54-68.

[13] THEODORIDIS S, KOUTROUMBAS K. 模式识别（第三版）[M]. 北京: 电子工业出版社, 2006.
THEODORIDIS S, KOUTROUMBAS K. Pattern Recognition(Third Edition)[M]. Beijing: Publishing House of Electronics Industry, 2006.

[14] NEEDLEMAN S B, WUNSCH C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of Molecular Biology, 1970, 48(3):443-453.

[15] SMITH T F, WATERMAN M S. Identification of common molecular subsequences[J]. Journal of Molecular Biology, 1981, 147(1):195-197.

[16] DAIN O, CUNNINGHAM R K. Building scenarios from a heterogeneous alert stream[A]. Proceedings of the 2001 IEEE Workshop on Information Assurance and Security[C]. West Point, NY, USA, 2001. 231-235.

[17] 郑宗汉, 郑晓明. 算法设计与分析[M]. 北京: 清华大学出版社, 2005.
ZHENG Z H, ZHENG X M. Design and Analysis of Algorithms[M]. Beijing: Tsinghua University Press, 2005.

[18] MIT Lincoln Lab. DARPA 2000 intrusion detection scenario specific dataset[EB/OL]. http://www.ll.mit.edu/IST/ideval/data/2000/2000_data_index.html, 2000.

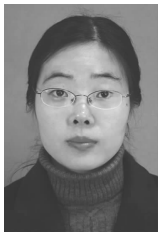
[19] NING P. TIAA: a toolkit for intrusion alert analysis[EB/OL]. http:// discovery. csc.ncsu.edu/ software/correlator/, 2009.

[20] NING P, CUI Y, REEVES D. Constructing attack scenarios through correlation of intrusion alerts[A]. Proceedings of the 9th ACM Conference on Computer and Communications Security[C]. Washington, D C, 2002. 245-254.

作者简介：

梅海彬（1973-），男，湖南常德人，东南大学博士生，主要研究方向为网络安全。

龚俭（1957-），男，上海人，博士，东南大学教授、博士生导师，主要研究方向为网络安全和网络行为学。



张明华（1977-），女，四川宣汉人，博士，上海海洋大学讲师，主要研究方向为普适计算和网络安全。