一种快速网络入侵检测的关联规则挖掘算法

丁　宏　赵观军
(杭州电子科技大学计算机学院,杭州　310018)
E-mail:zhaoguanj@126.com

摘　要　针对网络入侵检测领域使用关联规则挖掘关联模式精度不够,效率不高的问题。文章提出了一种新的基于最大值限制的关联规则算法,提出运用领域划分方法对特征项进行标准化处理,并结合各项的特征,给不同的特征项设置不同最小支持度,使挖掘模式更精确,挖掘速度明显提高。

关键词　入侵检测　关联规则　最小支持度　最大值限定

文章编号　1002-8331-(2006)11-0153-04　**文献标识码**　A　**中图分类号**　TP393

A Fast Association Rules Mining Algorithm for Network Intrusion Detection

Ding Hong　Zhao Guanjun
(College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018)

Abstract: In the network intrusion detection, association rules algorithm is used to extract relative rules, but its processing precision and efficiency are not satisfactory.In order to resolve this problem, this paper proposes a new association rules algorithm based on maximum constraint, proposes to use domain partition method to standardize each property item, and set different min-support related to their specialty, its precision and velocity are improved obviously.

Keywords: intrusion detection, association rules, minimum support, maximum constraint

1　引言

入侵检测(Intrusion Detection)作为一种动态防御技术,能抵御内外攻击,保护系统的保密性、完整性和有效性[1],已被广泛应用于网络安全的各个领域。入侵检测按检测方法可分为误用检测和异常检测,前者利用已知的攻击或系统漏洞来识别入侵,后者通过建立正常行为模式来检测入侵,它们的检测性能均取决于模型特征库的完备性。由于异常检测系统具有较好的系统独立性,通用性强,能检测未知的新型攻击,因此成为目前研究热点。

关联规则[2]是一项非常流行的技术,它是从一个数据库表中抽取多个特征的相关性。为了提高异常检测系统的性能,Lee和Stolfo[3]提出将基本的关联规则算法扩展到捕捉程序执行和用户活动相一致的行为,以此来构造正常行为模式库,然后根据当前系统活动来挖掘频繁出现的行为模式,将这个模式与正常的模式进行比较,以此来计算行为的差异性,一旦确定为入侵行为,就发出警报。

在以往处理关联规则时,项集中大都只设置单一最小支持度作为阈值,以此来获得满足条件的频繁项。但在实际应用中,尤其在网络入侵检测中,处理的审计数据属性很多,各属性的重要性不尽相同,考虑到系统的分析能力和准确性,需要给各属性项设置不同的最小支持度。本文针对目前网络入侵检测领域使用关联规则挖掘关联模式精度不够、效率不高的问题,提出了一种新的网络行为模式挖掘算法——基于最大值限制的关联规则算法,并运用领域划分方法对特征项进行标准化处理,然后结合各项的特征,给不同的特征项设置不同最小支持度,使挖掘模式更精确,挖掘速度明显提高。实验结果表明,提出的算法是有效的。

2　研究背景和相关工作

关联规则挖掘是一项非常重要的数据挖掘技术,从网络审计数据挖掘一条关联规则的实例是:ftp→get(0.1,0.4),它表示某用户有10%(支持度)的时间在使用ftp服务,其中有40%(信任度)的ftp服务用于get命令。将数据挖掘方法应用于审计数据,其目的是为了计算模型,它可以捕捉入侵和非入侵行为。Lee和Stolfo[3]运用关联规则和频集挖掘方法对系统审计数据进行挖掘,以此形成用户正常行为样式库,然后对所有的后续系统活动进行分析,来挖掘频繁出现的样式,将该样式集与正常样式库进行比较,然后采用相似度函数来计算差异度,任何重要的差异度将用来表示异常活动。

关联规则挖掘的标准算法是Apriori算法[2],他是由Agrawal等人于1993年提出的。他把挖掘过程分成两个阶段:第一阶段,通过扫描事务数据产生候选项集,如果这些项集中项的个数超过预定义的阈值(最小支持度),这个项集就被认为是大项集;第二阶段,从第一阶段产生的大项集中推出所有可能的关联规则,并把那些置信度大于预定阈值(最小置信度)的规则当作最终的输出。

之后,人们提出了各种基于Apriori的改进算法,这些算法大都针对某个具体领域或具体数据处理方式提出新的见解。我们知道对Apriori进行改进,不免对最小支持度和最小置信度进行调整。但前面很多方法大都停留于对单个最小支持度进行

基金项目:浙江省自然科学基金资助项目(编号:Y104426);浙江省教育厅高校科研计划资助项目(编号:20040457)
作者简介:丁宏,副教授,硕士生导师,主要研究方向为信息安全与信息处理;赵观军,硕士研究生,主要研究方向为网络与信息安全。

改进，其中 J.Luo[4]曾提出使用模糊算法来减小单个阈值带来的“边界差异”。在入侵检测中，审计数据各属性项特性不尽相同，需要用不同的标准来判断各项的重要性。所以 Liu 等人[5]提出一种算法用多个最小支持度来挖掘关联规则，他允许用户给不同的项指定不同的最小支持度，并把项集的最小的支持度定义为各子项的最小支持度的最小值。Wang 和 Han[6]在此基础上提出用各子项的最小支持度定义的函数来表示项集最小支持度。虽然这些算法设置最小支持度值的方式更加灵活，但跟普通算法相比，其算法复杂度较大。

3 提出的算法

考虑到网络入侵检测的特殊性，审计数据中各特征项拥有不同的重要性，需要给各项设置不同的最小支持度，但多个最小支持度必定给处理带来难度，所以提出“最小支持度最大值限制”来查找大项集，也就是用项集中各项的最小支持度的最大值来定义项集的最小支持度。我们把该关联规则算法定义为基于最大值限制的关联规则算法(Maximum Constraint Based Association Rules，简称 MCBAR)。该算法不仅比 Wang 等人[6]提出的函数方法简单，而且保留了层层迭代的处理方式，使算法始终保持向下收敛属性，并具有更好的挖掘效果。

本算法先通过比较预定义的最小支持度从事务中找到 1 个项目的大项集 L1，然后从 L1 中得到 2 个项目的候选项集 C2。注意，C2 中的所有候选项目的支持度必须大于等于 L1 中各项的最小支持度的最大值。这个特性在查找大项集前能起到很好的裁减作用。然后通过比较对应的最小支持度来查找得到 2 个项目的大项集 L2。这个过程一直继续，直到产生所有可能的大项集。

MCBAR 算法：

输入：一组包含 n 条网络审计记录的数据集(事务集)，一个由 p 个项目(某个属性项与对应的属性值) t_i 构成的项目集和每个项目预定义的最小支持度 $m_{t_i}(i=1,\cdots,p)$，以及一个最小置信度 λ。

输出：一组满足条件的关联规则。

步骤 1 计算网络审计数据集中各项目 t_i 出现的个数 c_{t_i}，计算其支持度 $s_{t_i}=\frac{c_{t_i}}{n}$。

步骤 2 根据支持度 s_{t_i} 生成大项集 L_1，即 $L_1=\{t_i|s_{t_i}\geq m_{t_i},1\leq i\leq p\}$。

步骤 3 设 r=1(r 表示当前项目集中项目的个数)，重复执行下列步骤，直至 $L_r=\Phi$。

步骤 3.1 由大项集 L_r 产生候选项集 $C_{r+1}=\{I_1,I_2,\cdots,I_k,\cdots\}$，各候选项 I_k 中的每一项目 x_i 满足：对于 $\forall x_i\in I_k$，有 $s_{x_i}\geq m_{I_k}=\max(m_{x_1},m_{x_2},\cdots,m_{x_i},\cdots)$，其中 m_{x_i} 为各项目 x_i 预定义的最小支持度。

步骤 3.2 计算 C_{r+1} 中各候选项 I_k 在网络审计数据集中的个数 c_{I_k}，以此得到各候选项 I_k 的支持度 $s_{I_k}=\frac{c_{I_k}}{n}$。

步骤 3.3 根据支持度 s_{I_k} 生成下一个大项集 L_{r+1}，即：$L_{r+1}=\{I_k|s_{I_k}\geq m_{I_k},1\leq k\leq|C_{r+1}|\}$。

步骤 3.4 r=r+1

步骤 4 根据生成的各大项集 $L_q(q\geq 2)$，为其中的每个大项 $I_k=\{I_{k_1},I_{k_2},\cdots,I_{k_q}\}$ 创建所有可能的关联规则：$R_k=I_{k_1}\wedge\cdots\wedge I_{k_j}\wedge I_{k_{j+1}}\wedge\cdots\wedge I_{k_q}\rightarrow I_{k_j}$，j=1，2，…，q。然后根据下列公式计算每条关联规则的置信度：

$$conf(R_j)=\frac{s_{I_k}}{s_{I_{k_1}\wedge\cdots\wedge I_{k_{j-1}}\wedge I_{k_{j+1}}\wedge\cdots\wedge I_{k_q}}} \quad (1)$$

步骤 5 将计算所得的置信度与预定义置信度 λ 比较，把大于等于 λ 的关联规则作为最终结果输出。

4 实验分析

为了验证 MCBAR 算法的有效性，我们使用 KDD CUP 99 作为实验数据来挖掘网络行为的关联规则。

4.1 数据源分析

KDD CUP 99 是美国林肯实验室模拟创建的网络数据包，它已成为国际社会公认的标准网络数据，每条数据有 41 个特征值，这些特征值可分为三大类：单个 TCP 连接的基本特征、使用 2s 时间窗口的流量特征和使用域知识的内容特征。为保证规则的合理性，我们根据 Lee 提出的轴属性(axis attribute)思想[3]，先对特征项进行筛选，因为那些不重要的属性项只会增加计算的复杂度，而且由这些属性生成的关联规则意义不大。考虑到实际网络状态是动态变化的，它会随着时间而改变，因此，对入侵检测而言，必须能识别新的异常情况，这就要求考虑时间窗口特征，它反应了网络连接记录的临时特征。而基于域知识的特征反应了网络流量的总情况，它能防止行为模式因局部范围变化而发生抖动。如 num_failed_logins，它表示登录失败的次数，所以 Stolfo 等也把它称为内容特征。基于以上分析，我们选择表 1 所示的 10 个特征属性作为挖掘项目。

表 1 筛选后的特征项

项目	特征属性	项目	特征属性	项目	特征属性
A	duration	B	protocol_type	C	service
D	flag	E	src_bytes	F	dst_bytes
G	num_failed_logins	H	logged_in	I	srv_count
J	serror_rate				

4.2 特征项标准化

在实验前，从 KDD CUP99 获得 5 000 条形如表 2 的正常网络审计数据：

表 2 KDD CUP 99 测试数据

A	B	C	D	E	F	G	H	I	J
0	udp	private	SF	105	146	0	0	1	0
0	udp	private	SF	105	146	0	0	1	0
0	udp	domain_u	SF	29	0	0	0	1	0
0	tcp	http	SF	223	185	0	1	4	0
0	tcp	http	SF	272	1 861	0	1	18	0.06
1	tcp	smtp	SF	3 170	329	0	1	2	0
…	…	…	…	…	…	…	…	…	…

从表中可以看出，各特征项的取值范围有所不同，有的是“离散”的，如 B(protocol_type)，其值只由 tcp、udp 等少数几种协议类型组成。而有的则是“连续”的，如 D(src_bytes)，它可以是网络协议限定的最大报文段长度(MSS)以内的任意值。对于连续项目，由于其值分布的广泛性，使得某些小概率项对寻找频繁项没有帮助。因此在使用 MCBAR 算法前应先对连续型的属性项进行“离散化”处理，这个过程有点类似于聚类方法，即根据属性项值的分布特性，找出数据分布最集中的几个聚集点，称为“核点”(Kernel Point)，核点数值对原始连续项最具代

表性。然后利用核点将连续数据分成若干个领域，即以核点为中心，对各核点之间的区域进行划分。例如，存在两个核点 A 和 B，且存在某个点 e 位于 A，B 之间（即 A<e<B），如果 e 到 A 的距离比到 B 的距离近，则认为 t 属于 A 领域；反之，t 属于 B 领域。但是对于最右的边界核点，我们以该核点为中心，把其左右领域设置成相等长度，而把剩余的区域单独作为一个领域。但是有一种情况例外，当区域中只存在一个核点时，我们把区域只划分成两个领域，其中一个领域就是核点，剩余部分形成另一个领域。这样，每一个连续项都能以核点为中心，划分成多个领域，我们把这些领域称为标准化的项。

例如，实验中属性项 E（src_bytes）值的统计分布特性如图 1 所示。很显然，E=0，E=105，E=520，E=1 032 四点的分布最为集中，它们成为属性项 E 的四个核点，根据这四个核点，属性项 E 可分为 E1（0≤E≤52），E2（52<E≤312），E3（312<E≤776），E4（776<E≤1 288），E5（E>1 288）五个领域，该五个领域就是 5 个与属性项 E 有关的标准项。但是在运用到关联规则挖掘时必须注意，由相同特征项划分得到的标准项不能同时出现在多项集的某个项中，即组成多项集的项必须来自不同的特征属性。

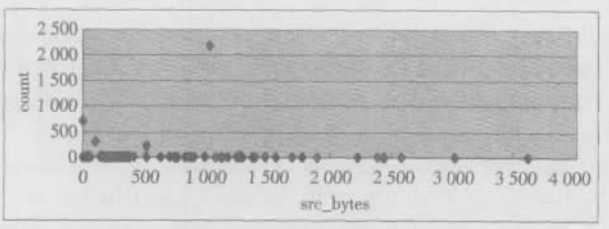

图 1　属性项 E 统计分布特性

4.3　实验结果

经过前面的标准化处理，所有的连续项都能划分为由域组成的标准项，这样就得到了 MCBAR 算法处理的项集，根据具体情况设置各特征项的初始最小支持度如表 3。

表 3　各特征项初始设置的最小支持度

Item	A=0	A=1	B=tcp	B=udp	C=http	D=SF	E1	E2	E3	E4	E5	…
Min-sup	0.3	0.05	0.3	0.2	0.3	0.3	0.2	0.1	0.1	0.3	0.1	…

根据步骤 1 可以从审计数据中计算得到各项的支持度如表 4。

表 4　各特征项计算得到的最小支持度

Item	A=0	A=1	B=tcp	B=udp	C=http	D=SF	E1	E2	E3	E4	E5	…
Min-sup	0.97	0.029	0.669	0.327	0.592	0.679	0.204	0.147	0.074	0.560	0.018	…

根据步骤 2 得到大项集 L1 为：{{A=0}，{B=tcp}，{B=udp}，{C=http}，{D=SF}，{E1}，{E2}，{E4}，…}。

根据步骤 3.1，要得到候选项集 C2，要求 2-项集的成员的支持度必须大于或等于前面预定义的最小支持度的最大值。例如{A=0，B=tcp}是满足条件的候选 2-项集成员，因为 A=0 的支持度（0.97）和 B=tcp 的支持度（0.669）都大于 0.5。而{B=tcp，E2}则不行，因为 E2 的支持度（0.147）小于 B=tcp 的支持度（0.4）。根据这个原则，得到候选 2-项集 C2 如下：

{{A=0，B=tcp}，{A=0，C=http}，{A=0，D=SF}，{A=0，E4}，{B=tcp，C=http}，{B=tcp，D=SF}，{B=tcp，E4}，{B=udp，D=SF}，{B=udp，E4}，{C=http，D=SF}，{C=http，E4}，{D=SF，E4}…}

根据步骤 3.2，计算各候选 2-项集各项的支持度，即各项在总记录中出现的概率，如下：

sup（A=0，B=tcp）=0.641　　sup（A=0，C=http）=0.589
sup（A=0，D=SF）=0.970　　sup（A=0，E4）=0.559
sup（B=tcp，C=http）=0.592　　sup（B=tcp，D=SF）=0.668
sup（B=tcp，E4）=0.003　　sup（B=udp，D=SF）=0.327
sup（B=udp，E4）=0　　sup（C=http，D=SF）=0.592
sup（C=http，E4）=0.090　　sup（D=SF，E4）=0.560

根据步骤 3.3，把上述支持度与对应子项的最小支持度的最大值相比较，得到大项集 L2：

{{A=0，B=tcp}，{A=0，C=http}，{A=0，D=SF}，{A=0，E4}，{B=tcp，C=http}，{B=tcp，D=SF}，{B=udp，D=SF}，{C=http，D=SF}，{D=SF，E4}…}

如此继续，我们可以得到候选项集 C3，再得到大项集 L3，直到不能产生新的大项集为止，这时就得到了满意的关联规则集。从前面列举的情况可以发现，用最大值限制关联规则算法裁减效果很好，在从候选项集 C2 中挖掘大项集 L2 时，该算法快速消除了大部分包含 src_bytes 的无效项，符合挖掘有效频繁项的目标。比如{B=tcp，E3}，该关联规则对检测入侵是没有意义的，提前排除有利于后面挖掘有效的规则。

实验中，我们设置最小置信度 λ=0.8，并利用公式（1）计算各关联规则的置信度，得到{A=0，D=SF，E4}（0.559，0.691），{A=0，B=tcp，C=http，D=SF}（0.329，0.970）的关联规则。由于前一条关联规则的置信度小于 λ，因而只保留后面的 4 项关联规则。该 4 项关联规则表示一个正常的网络连接，它的延时为 0，协议类型为 tcp，服务类型为 http，协议标志为 SF，它的支持度为 0.329，置信度为 0.970。

4.4　算法的分析和比较

与以往算法相比，MCBAR 算法的最大特点是不管生成候选项集还是大项集，都使用各子项集最小支持度的最大值作为限制门限。从上面的实验结果可以看出，2-候选项集个数只占 1-大项集组合数的 50%，而如果采用 Liu[6]提出的 MMSAR 算法来挖掘关联规则，产生的候选 2-项集个数却等于 1-大项集组合数，没有任何收敛，在后面的层层迭代的过程中，收敛效果也没有 MCBAR 算法好。候选项集的快速缩减，大大减少了后面大项集的产生个数，同时也减少了处理时间，提高了挖掘效率。而与普通 Apriori 算法相比，它允许给不同特征项设置不同的最小支持度，使算法侧重于重要特征项挖掘，有效减少了无效特征项的干扰。

前面我们只对 KDD 小部分数据进行测试，若对包含不同记录数的审计数据包进行实验，其实验结果如图 2 所示。

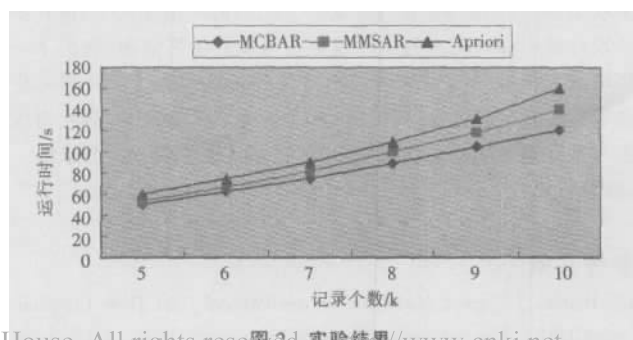

图 2　实验结果

从图中可以发现，MCBAR 算法在执行效率上比 Apriori 算法和 MMSAR 算法有了一定的提高，而且随着记录个数的增加，效果更加显著。

综上可得，在入侵检测领域，基于最大值限制的关联算法具有比普通 Apriori 算法更好的挖掘效果。通过给不同的特征项设置不同最小支持度，使挖掘模式更精确，更能反应网络的正常行为。

5 结论

本文把最大值限制关联规则技术运用到网络异常入侵检测中，根据 MCBAR 算法的特点对连续型特征项进行了标准化处理。实验证明 MCBAR 算法在挖掘数据过程中具有更好的裁减效果，它能提高网络入侵检测的效率，给不同特征项设置不同最小支持度，使挖掘的模式更精确。本文使用领域划分的方法使特征项得到了标准化，但是给各特征项设置不同的最小支持度仍是从经验出发，所以今后研究重点是寻找更好的方法，使各特征项最小支持度的设置更智能化。

（收稿日期：2005 年 8 月）

参考文献

1.R Heady, G Luger, A Maccabe et al.The architecture of a network level intrusion detection system.Computer Science Department, University of New Mexico, 1990-08

2.Agrawal R, Imielinski T, Swami A.Mining association rules between sets of items in large databases.SIGMOD-1993, 1993: 207~216

3.Wenke Lee, Salvatore J Stolfo.Data Mining Approaches for Intrusion Detection[C].In: Proceedings of the 7th USENIX Security Symposium San Antonio, Texas, 1998-01: 79~94

4.Jianxiong Luo.Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection[J].International Journal of Intelligent Systems, 2000; 15(1)

5.Bing Liu, Wynne Hsu, Yiming Ma.Mining Association Rules with Multiple Minimum Supports[C].In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(KDD-99), San Diego, CA, USA, 1999-08

6.K Wang, J Han.Mining frequent itemsets using support constraints[C].In: Proceedings of the 26th 374 International Conference on Very Large Data Bases, 2000: 43~52

（上接 23 页）

表 1 约简协议实例

SRP2-1	SRP3-1	SRP3-2
$A\rightarrow B:\{A, N_A\}_{PubB}$	$A\rightarrow B:\{B, N_A\}_{K_{AS}}$	$A\rightarrow B:N_A$
$B\rightarrow A:\{B, N_A, N_B\}_{PubA}$	$B\rightarrow S:\{A, N_B\}_{K_{BS}}, \{B, N_A\}_{K_{AS}}$	$B\rightarrow S:\{N_A, A, N_B\}_{K_{BS}}$
$A\rightarrow B:\{N_B\}_{PubB}$	$S\rightarrow A:\{A, B, N_A\}_{K_{AS}}, \{A, N_A, N_B\}_{K_{BS}}$	$S\rightarrow A:\{B, N_A, N_B\}_{K_{AS}}$
	$A\rightarrow B:\{A, N_A, N_B\}_{K_{BS}}$	$A\rightarrow B:N_B$

利用构造冗余协议，再使用约简检测有效地降低使用散乱的信息直接构造协议而产生的可能协议的空间。

由于采用了基于认证检测的全信息项构造冗余协议，我们构造的协议满足文献[2]中提出的错误-停止协议，但文献[2]中的分析将 NS 协议看作是正确的错误-停止协议，这个问题产生的原因正是因为协议交换的消息项（协议的第二条消息）不是全信息的。在对协议的设计过程中，我们也得到了一些关于协议规范的安全性的经验：协议规范除了需要指明的所交换的消息以及在参与方交换消息的步骤，其中指明协议消息中各个消息的类型；协议规范需要指明消息交互的动作，包括消息的顺序，以及接收方在发送消息之前对接收的消息进行验证执行的动作，以真正实现错误-停止体制，如 CCITTX.509 协议正是因为协议规范的描述有缺陷而导致漏洞的存在。我们在协议中加入了效率的约简规则。

我们使用的方法仍然是启发式方法，如使用认证检测构成协议组件。相应的，随着对协议的分析的进一步深入，可以在协议设计器中加入更多的组件或更多可能的约简规则。我们下一步的研究内容是考虑如何增加协议分析器和协议设计器分析和设计较复杂的协议，如群组协议，这类协议具有动态性，协议参与者随着时间的变化而变化，这增加了设计和分析的难度。

（收稿日期：2006 年 1 月）

参考文献

1.N Heintze, J Tygar.A Model for Secure Protocols and Their Compositions[J].IEEE Transactions on Software Engineering, 1994; 22(1): 16~30

2.L Gong, P Syverson.Fail-Stop protocols: an approach to design secure protocols[C].In: Proceedings of DCCA-5 Fifth International Working Conference on Dependable Computing for Critical Applications, Oakland: IEEE Computer Society Press, 1998: 79~100

3.C Meadows.Formal verification of cryptographic protocols: a survey[C].In: Proceedings of ASIACRYPT'94, Springer Verlag, 1995: 135~150

4.C Rudolph.A Formal Model for Systematic Design of Key Establishment Protocols[C].In: Proceedings of the Third Australasian Conference on Information Security and Privacy, Brisbane Queensland, Australia: Springer Verlag, 1998: 332~343

5.L Buttyan, S Staamann, U Wilhelm.A simple logic for authentication protocol design[C].In: Proceedings of the IEEE Computer Security Foundations Workshop XI, Rockport, USA: IEEE Computer Society Press, 1998: 153~162

6.A Perrig, D X Song.A first step towards the automatic generation of security protocols[C].In: Proceeding of Network and Distributed System Security Symposium, 2000-02

7.S Perrig, D X Song.Looking for diamonds in the desert: Extending automatic protocol generation to three-party authentication and key agreement protocols[C].In: Proceedings of the 13th IEEE Computer Security Foundations Workshop, IEEE Computer Press, 2000

8.Joshua D Guttman.Protocol design via the authentication tests[C].In: Proceedings of 15th IEEE Computer Security Foundations Workshop, IEEE Computer Society Press, 2002: 92~103

9.F J Thayer, J C Herzog, J Guttman. Strand spaces: Why is a security protocol correct[C].In: Proceedings of the 1998 IEEE Symposium on Security and Privacy, Los Alamitos: IEEE Computer Society Press, 1998: 160~171

10.John A Clark, Jeremy L Jacob.Protocols are Programs Too: the Meta-heuristic Search for Security protocols[J].Information and Software Technology, 2001; 43: 891~904

11.王张宜，李莉，张焕国.网络安全协议的自动化设计策略[J].计算机工程与应用，2004; 41(5): 16~17

12.李莉.安全协议的形式化分析及验证技术[D].博士学位论文.武汉大学，2004