基于信息增益的防火墙过滤域排序优化①

Optimizing Sequence of Filtering Domain of Firewall Based on Information Gain

王卫平 王旭虓 陈赫然 陈家耀

(中国科学技术大学 信息管理与决策科学系 安徽 合肥 230026)

摘 要: 传统的防火墙从检测规则冲突和调整规则排序两方面来提高防火墙性能，但效果都不是很理想。本文从优化防火墙过滤域排序这样的一个新角度，依据信息增益理论构造决策树，然后根据决策树层次与防火墙过滤域排序的对应关系，确定了过滤域的最优排序。实验表明，这个最优排序确实能够显著地降低数据包与防火墙规则的元组比较的总次数，从而提高防火墙的过滤效率。

关键词: 防火墙 信息增益 决策树 过滤域排序 元组比较

1 引言

自计算机网络问世以来，网络的安全性一直是人们讨论的话题。随着网络的迅速发展和广泛使用，人们在得益于网络的同时，网上的数据也受到了黑客的攻击，所以网络安全变得越来越重要。当资源共享广泛用于政治、军事、经济以及科学的各个领域，网络的用户来自社会的各个阶层与部门，大量在网络中存储和传输的数据都有可能被盗用和篡改。人们为了保护数据和资源的安全，创建了防火墙。

2 防火墙

防火墙技术是当今网络信息安全的核心技术之一，也是抵御外部网络攻击的第一道屏障。防火墙是置于内部可信网络和外部不可信网络之间的一个安全组件，用来控制跨越网段的流量。从技术角度来看，防火墙是作为一种连接内部网络和外部网络的网关，提供对进入内部网络连接的访问控制能力。

防火墙能根据预先定义的防火墙安全策略，对流入流出的数据包进行逐一检查，允许合法连接进入内部网，阻止非法连接，从而保证内部网络的安全。这样，防火墙安全策略的配置就显得异常重要，一个好的配置策略通常能大大提升防火墙过滤效率。

防火墙的安全策略实际上是一个访问控制列表。表 1 即是一个防火墙访问控制列表的实例。

表 1 防火墙访问控制列表实例

规则号	协议	源地址	源端口	目的地址	目的端口	行为
1	TCP	192.168.0.6	any	*.*.*.*	80	deny
2	TCP	192.168.0.*	any	*.*.*.*	90	accept
3	UDP	*.*.*.*	any	202.38.64.*	53	accept
4	UDP	*.*.*.*	any	*.*.*.*	any	deny

表中的访问控制列表由很多列表项组成，每个列表项称为一条规则，也就是防火墙规则。每条规则都是由三部分组成：规则号、过滤域和动作域。规则号决定过滤规则排列的先后顺序；过滤域是规则的主体部分，可以有很多项构成，但是常用的有 5 项：协议、源 IP 地址、源端口、目的 IP 地址和目的端口；动作域包括接受和拒绝。

当数据流进入防火墙后，防火墙检查数据包的有关头信息(IP 地址、端口号等)，然后将这些头信息与防火墙的第一条规则中对应的内容逐一匹配，如果匹配成功则按照规则的动作域内容进行操作，而不再匹配后续的规则，否则继续与下一条规则匹配，直到匹

① 收稿时间:2008-11-13

配成功为止。这一比较的过程就称为元组比较，元组比较的次数决定了防火墙过滤数据的时间，这也是衡量防火墙效率的一个标准。

3 相关的研究工作

防火墙一般位于内外部网络的边界，用来实现对进入内部网络连接的访问控制。但正是由于防火墙是内外网之间的交通要道，要检查所有内外网之间相互交换的数据包，当网络数据流量较大的时候，防火墙就会称为整个内部网络和外界通信的瓶颈，导致网络速度减慢甚至于瘫痪。因此，增强防火墙的过滤能力，提高防火墙的性能一直都是防火墙研究领域中的重要内容。

国内外关于防火墙的研究主要集中于两大方面：

(1) 防火墙规则间的冲突检测：这是与安全直接相关的。防火墙规则在设定、添加、删除和修改的过程中，都有可能引发规则间的冲突，带来安全隐患。为了解决这一问题，文献[1]通过建立 CSP(Constraint Satisfaction Problem)去侦测安全策略和规则集合间的矛盾；文献[2]提出一种基于元组空间搜索的规则集冲突检测算法，时间复杂度低，速度快，但不适用于分布式规则集冲突检测；文献[3]针对分布式防火墙，探讨了规则间的可能关系和错误的策略配置，最后给出检测错误的算法。

(2) 防火墙规则设计及性能提高：好的规则设计不仅能减少规则间的冲突，更能够优化防火墙的过滤性能。防火墙性能可以通过软件和硬件手段来提高，软件方面主要体现在对规则的调整和排序上，文献[4]提出剔除防火墙中异常规则，合并相类似规则，可以优化规则；文献[5]将规则映射成多维空间中的一个几何体，图形化处理后，再将几何体映射成规则；文献[6]提出基于统计分析，对数据包统计分析，然后调整规则的顺序，以减少匹配时间。

由于传统的防火墙技术对规则表的过滤域属性排序是没有要求的，本文试图从一个新的角度看问题，我们将运用信息增益理论对过滤域中的各个属性列进行分析，从而找到一种最优排列顺序，使得数据包和防火墙规则的总元组比较次数最少。

4 信息增益

信息增益是信息论中一个重要的概念，广泛应用于机器学习和数据挖掘领域，使用信息增益来评价属性在分类过程中所体现的信息量的程度。

信息增益是指期望信息或者信息熵的有效减少量，根据它能够确定在什么样的层次上选择什么样的变量来分类。若属性的信息增益越大，则在分类过程中对类的划分也将起到较大的作用，因此，在进行属性选取的过程中，通常选取信息增益值大的属性作为分类的依据。信息增益的计算如下[7]：

① 样本的分类期望信息

假设 S 是 s 个数据样本的集合。假定类标号属性具有 m 个不同值，定义 m 个不同类 $C_i(i=1,2,\cdots,m)$。设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息为：

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

其中 p_i 是任意样本属于 C_i 的概率，并用 s_i/s 估计。

② 属性 A 划分为子集的熵

设属性 A 具有 v 个不同值$\{a_1,a_2,\cdots\cdots,a_v\}$。可以用属性 A 将 S 划分为 v 个子集$\{S_1,S_2,\cdots\cdots,S_v\}$；其中 S_j 包含 S 中这样一些样本，它们在 A 上具有值 a_j。如果 A 选作测试属性(即最好的分裂属性)，则这些子集对应于由包含集合 S 的节点生长出来的分枝。设 s_{ij} 是子集 S_j 中类 C_j 的样本数。根据 A 划分成子集的熵或期望信息为：

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \quad (2)$$

其中，项 $\frac{s_{1j} + \ldots + s_{mj}}{s}$ 充当第 j 个子集的权，并且等于子集中的样本个数除以 S 中的样本总数。

③ 信息增益

$$Gain(A) = I(S_1, S_2, \ldots, S_m) - E(A) \quad (3)$$

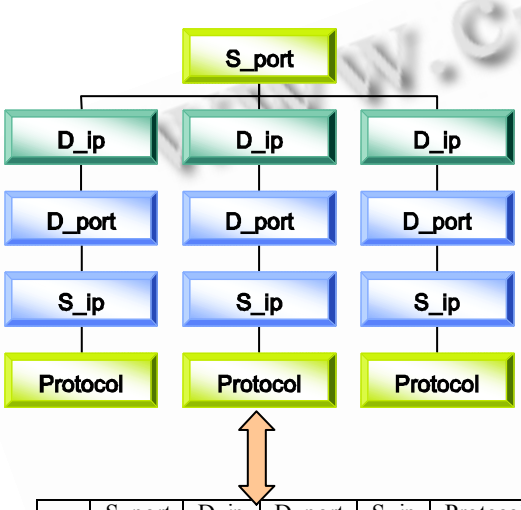
因此，Gain(A)是由于知道属性 A 的值而导致的熵的期望压缩。

5 过滤域排序优化

防火墙系统本身就是一种包分类系统[8]，数据包进入防火墙后，从防火墙的第一条规则开始，与过滤域中的每一项逐一匹配。若数据包与规则不匹配则说明数据包和这条规则不属于同一类；若数据包与规则完全匹配，那么数据包和这条规则就是同一类的，我

们可以根据这个分类来接收或拒绝数据包，因此，数据包与规则匹配的过程就是对数据包的一个分类过程。所以我们完全可以将分类技术中的信息增益理论应用于防火墙规则匹配领域中，通过优化防火墙过滤域属性排序来减少数据包与规则的总元组比较的次数，缩短比较时间。

防火墙的规则表就是一个训练集，我们可以根据预测进入防火墙的数据来对每条规则设定一定的概率。由于过滤域属性中常用的有 5 项：源 IP 地址，目的 IP 地址，源端口，目标端口和协议，那么利用信息增益理论可以构造出一棵决策树，这棵决策树的层结构与防火墙的过滤域排序是对应的，层结构的属性序列也即是过滤域的最优排序，如图 1 所示。



	S_port	D_ip	D_port	S_ip	Protocol
1					
2					
3					
4					
5					
…					

图 1　决策树层次与过滤域排序的对应关系

构造决策树的算法是递归算法，算法如下：

① 分别对过滤域属性利用公式(3)计算信息增益，并取信息增益最大的属性作为分裂属性 A，即决策树的根结点(图 1 中的 S_port)；

② 将属性 A 取值相同的规则归于同一子集 A_i，一个子集对应一个分支；

③ 计算 A_i 中规则数占所有规则的权重 W_i；

④ 对 A_i 中剩下的属性利用公式(3)计算信息增益，取信息增益最大的属性作为可能的分裂属性 Z；

⑤ 对相同的属性 Z 对应的权重 Wi 求和∑Wi；

⑥ 选择∑Wi 最大的属性为第二分裂属性 B，即决策树第二层的所有节点都必须为 B(图 1 中的 D_ip)；

⑦ 若子集内规则的所有属性均已测试，则结束；否则，重复②-⑥。

将防火墙的每条规则看作一个分类，将过滤域中的属性看作决策树的测试属性，那么就可以从规则表得到相应的决策树。从根到叶节点的一条路径就对应于一条防火墙规则，整个决策树就表示了一组防火墙规则集合。

需要说明的是：利用上述算法和利用 ID3 算法构造决策树是有区别的，ID3 算法构造的决策树每一层的属性可以不一样，而这里构造的决策树的每一层属性必须一样，因为防火墙过滤域排序的每一列必须是同一属性。

6　仿真实验

为了检验排序优化算法的正确性和执行效率，我们对算法进行了仿真，采用数据库来存储防火墙规则表，同时采用 C++语言在 CPU P4 1.7G、RAM 512M、Windows XP 操作系统的机器上实现了算法。

我们对天网防火墙和 look'n' stop 防火墙规则表做了适当修改，来构造我们的实验防火墙规则表，规则表有 5 个过滤域：协议，源 IP 地址，源端口，目标 IP 地址和目标端口。当规则数分别等于 40，80，120 时，由一个 100 条数据的数据集通过防火墙，计算在不同的过滤域排序下，这个数据集和防火墙规则表的总元组比较次数。

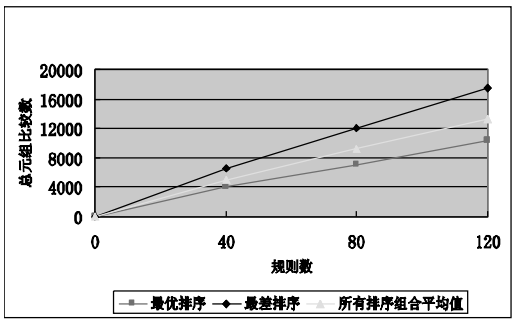


图 2　不同排序下的总的元组比较数

图 2 是当不同数目的规则时，按上述算法构造决策树，确定过滤域最优排序下数据包和规则元组比较的总数和最差排序下的总比较元组数，以及所有可能

的排序组合下比较元组数的平均值。从图中可以看出最优排序的总元组比较数确实要比最差排序的总元组比较数和所有排序组合的均值明显少很多，并且随着规则数目的增加，差值也在逐渐增加。实验证明，上述的算法得出的最优排序确实能显著的降低总的元组比较次数。

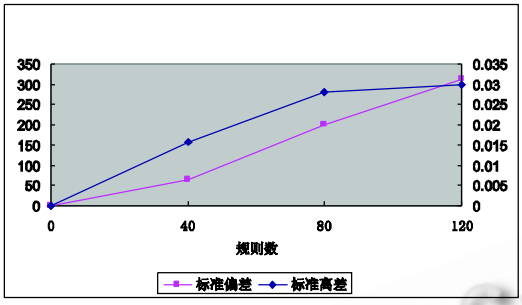


图 3　过滤域前两列确定后总元组比较数的标准偏差与标准离差

图 3 是过滤域前两列确定后，$P_3^3=6$ 种排列组合的总元组比较数的标准偏差与标准离差。利用上述算法构造两层决策树，确定了过滤域排序的前两列，从图 3 可以看到即使规则数达到 120，6 种排列的总元组比较数的标准偏差只有 313，标准离差(=标准偏差/期望)仅 0.03。显然，这几种排列下的总元组比较数非常接近。因此，当规则总数很大时，完整的构造决策树将会很复杂，那么我们不妨就只确定过滤域的前两列，这样只牺牲少许的精度，来换取较大的代价。

7　结论和展望

防火墙技术是企业网络安全的基础，研究人员已经做了大量的工作。本文从优化防火墙过滤域排序这样一个新的角度，来优化防火墙的安全配置策略，从而避免了对防火墙规则间关系的改动，更加不会引起规则间的冲突。

本文抓住防火墙实质上就是一个包分类系统，从信息增益理论出发构造出决策树，根据决策树层次与防火墙过滤域排序的对应关系，确定了过滤域的最优排序。仿真实验证明，这个最优排序确实能够显著地降低数据包与防火墙规则的元组比较的总次数，进而大幅度提高防火墙的过滤效率。而当防火墙规则数很多的时候，我们可以只确定过滤域排序的前两列，节省构造决策树的时间，并且不会影响防火墙的过滤能力。

当然，我们现在研究的只是静态的防火墙，首先要预测进入防火墙的数据，对防火墙的规则赋一定的权重，然后计算信息增益，构造决策树，确定最优过滤域排序。以后，我们还将致力于研究动态的防火墙，当外界数据变化时，到底要不要调整防火墙的过滤域排序，以及什么情况下，什么时候调整等等问题。

参考文献

1 Pozo S, Ceballo R, Gasca RM. CSP-Based Firewall Rule Set Diagnosis using Security Policies. Second International Conference on Availability, Reliability and Security, 2007: 723－729.

2 李林,卢显良.基于元组空间搜索的规则集冲突检测算法.北京邮电大学学报, 2006,29(5):111－114,124.

3 王卫平,陈文惠,朱卫未,陈华平,杨杰.分布式防火墙策略配置错误的分析与检测.中国科学院研究生院学报, 2007,24(2):257－265.

4 Katic T, Pale P. Optimization of Firewall Rules. Proceedings of the ITI 2007 29th Int. Conf. on Information Thechnology Interfaces, 2007.

5 张翼,张勇,汪为农.防火墙过滤规则的建模和全面优化.计算机工程与应用, 2006,42(6):146－150.

6 任安西,杨寿保,李宏伟.一种基于统计分析的防火墙规则匹配优化方法.计算机工程与应用, 2006,42(4):162－164.

7 Han JW, Kamber M. Data Mining: Concepts and Techniques.Beijing:China Machine Press, 2001:286－287.

8 王卫平,陈文惠,李哲鹏,陈华平.防火墙策略不一致性检测算法.中国科学院研究生院学报,2007,24(3):378.