一种基于个性化邮件特征的反垃圾邮件系统

鲁晓南,接 标
(安徽师范大学,安徽 芜湖 241000)

摘 要:垃圾邮件的内容因人而异,现有的垃圾邮件过滤系统大多采用统一的过滤标准对用户的邮件进行过滤,因而忽略了垃圾邮件的这种个性化特征。针对这一情况提出一种个性化垃圾邮件过滤的计算模型,它事先不需要对模型进行针对性的训练,从对用户日常处理不同类型邮件的行为中分析和挖掘垃圾邮件的个性化特征,然后利用这种个性化特征在对垃圾邮件进行识别的同时不断强化这种个性化特征,以实现逐步提升对垃圾邮件识别率的目的。据此实现了相应的原型系统,通过对此系统的实验验证,该方法在现实环境下对垃圾邮件具有很好的过滤效果。

关键词:垃圾邮件;模式识别;机器学习
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2009)08-0155-04

An Individual Anti-Spam Technology

LU Xiao-nan, JIE Biao
(Anhui Normal University, Wuhu 241000, China)

Abstract: The content of spam is different for different ones. Current anti-spam systems mostly employ a uniform filter standard to identify spam, thus they ignore this individuality of spam. As for it, puts forward an individual anti-spam model, which needn't train the system in advance. It analyzes and digs individualities of spam by the operations of user's processing different types of daily e-mails, then the individualities of spam is used to identify spams, meanwhile, the individualities of spam can be strengthened in the process. By this way it is achieved that identified ratio of spam is improved gradually. Based on it prototype system is completed, experiments on it prove that performance of the model for filterring spam is good in real environment.

Key words: spam; pattern recognition; machine learning

0 引 言

随着 Internet 的普及,电子邮件日益得到了广泛的应用,成为日常生活中人与人之间通信、交流的重要手段。但是随之而来的垃圾邮件也越来越猖獗。垃圾邮件对整个信息社会造成极大危害,例如浪费电子邮件用户的时间;占用网络、系统资源;对网络安全形成威胁等。面对垃圾邮件越来越泛滥的情况,反垃圾邮件技术目前已成为国内外研究的一个热点,目前主流的技术有"黑/白名单"、简单 DNS 检查、Bayesian 过滤器[1,2]等。

反垃圾邮件技术发展趋势是要针对垃圾邮件的变化也能动态灵活多变,能够自适应学习,更加智能化,更具有个性化。为此,提出一种全新的反垃圾邮件技术——一种基于个性化邮件信息的反垃圾邮件系统。

收稿日期:2008-11-21;**修回日期**:2009-02-23
基金项目:安徽师范大学青年科学基金项目(2007xqn56)
作者简介:鲁晓南(1978-),男,山东临沂人,硕士研究生,研究方向为搜索引擎、垃圾邮件处理、并行分布式计算、人工智能。

1 系统的主要特点

由于目前的大多数反垃圾邮件技术主要针对垃圾邮件的通用模式的进行研究和分析,然而垃圾邮件本身具有很强的个性化特点,一封邮件对一些人来说是垃圾邮件,但对另一些人来说可能并不是垃圾邮件。每个人的个性需求不同,相应每个人的垃圾邮件模式也是不同的,因而对垃圾邮件模式的研究不能过于统一化,一概而论。同理正常邮件模式也具有这种个性化特征,但目前的反垃圾邮件研究很少涉足正常邮件模式的研究。对此文中建立了这种垃圾邮件过滤系统,它不但对个性化正常邮件模式进行分析和挖掘,同时也对个性化的垃圾邮件模式进行分析和挖掘,将以上两者结合起来,来对用户的邮件进行分析和处理,因而它能够极大降低对垃圾邮件的漏识率和错识率,提高垃圾邮件过滤的效率。

另外此系统模型与目前大多数反垃圾邮件模型[3~8]不同的是,它不需要在正常使用之前对系统使用样本实例进行训练,用户只管进行正常的收发邮件操作,在开始时只是对日常所收到的邮件进行垃圾邮

件和正常邮件分类，系统模型自动对分类邮件进行分析处理，逐渐建立起正确的针对个人的个性化邮件模式，用它来实现对垃圾邮件的过滤，即此处它潜在地借用了人的个性化智能分类邮件的知识。因而此系统模型具有很强的自适应学习功能，当用户的邮件的个性化选择发生变化时，用户只要向系统中添加反映这种变化的邮件类，系统就会自动调整相应的模式以适应用户对邮件喜好的变化，同时依靠这种运作，能够快速学习识别新出现的垃圾邮件类和正常邮件类，这样真正实现了对邮件的个性化智能过滤。

2 系统的设计思想

由于在邮件内容中，各语句中的主干词(一般指主语与宾语等)，即这些主题词集合一般反映了此邮件所涉及的内容信息，它们大体能反映用户的兴趣范围，蕴含着用户的一种个性化信息模式。并且反复出现次数越多的主题词，越说明邮件的侧重内容，因此一封邮件对个人来说是否是垃圾邮件可以通过对这些主题词的判断来决定，这也是人工识别垃圾邮件常用的一种方法。另外一个人日常的邮件通讯一般局限在与自己有关的一些内容信息，在一定时期内它们会集中在某一些固定主题上，因而内容信息不会杂乱无章，毫无范围限制。

基于以上事实，从个人最初人为分类处理产生的正常邮件和垃圾邮件中分别提取出反映邮件内容的主题词集合，将每个邮件的这些主题词集合作为用户的一条个性化邮件模式，分别存放在正常邮件模式库与垃圾邮件模式库中，这样经过短时间的积累，就可以建好反映用户个性的正常邮件模式库和垃圾邮件模式库。在这之后从用户日常收到的各种邮件中提取出反映内容的主题词集合，将它们与正常邮件模式库和垃圾邮件模式库进行一系列复杂的匹配计算，然后对以上两类模式库的计算结果进行一些对比计算处理，最后从最终的计算结果中可以判断邮件是正常邮件还是垃圾邮件。

3 系统模型的具体设计

3.1 总体框架

图 1 为系统总体的工作框架图，它主要由常用主题词库、正常邮件模式库、垃圾邮件模式库等几部分组成，基本的工作流程是这样的：首先初始时用户手工分类选择一些邮件生成用户分类邮件库(包含正常邮件类与垃圾邮件类)，参照常用主题词库提取邮件中的主题词集合建造正常邮件模式库与垃圾邮件模式库。然后利用常用主题词库提取每封日常邮件中的主题词集合，将它们与正常邮件模式库和垃圾邮件模式库进行匹配计算，这之后再将这两类计算结果进行对比计算，最终由这个结果判断检测的邮件是否是垃圾邮件。

3.2 常用主题词库的建造

识别垃圾邮件的第一步就是将待查邮件内容中的主题词提取出来，而此提取过程就要用到常用主题词库，即通过用常用主题词库与邮件文本内容进行对照匹配来实现这一目的，同样建立邮件模式库也用到这一操作。常用主题词库中所含主题词主要是各种用户日常电子邮件通讯中常用到的词，并且这些词对邮件整体内容具有一定的影响作用。由于许多常用词，例如冠词、介词、连词等，经常在邮件文本中频繁出现，它们一般对邮件整体内容影响不大，因此这样的一类词不被包含在邮件主题词库中。

另外为了提取邮件文本中的主题词操作的方便，此系统也把一些主题词的变形形式(例如英文单词中的时态和数的变化形式等)和它们的词干形式都包含在邮件主题词库中。目前此系统模型主要针对中英文邮件文本信息进行处理，因而分别建立了中文主题词库与英文主题词库。为了提高建造邮件主题词库的效率，利用了目前互连网上公开的一些语料库，例如 Lin-

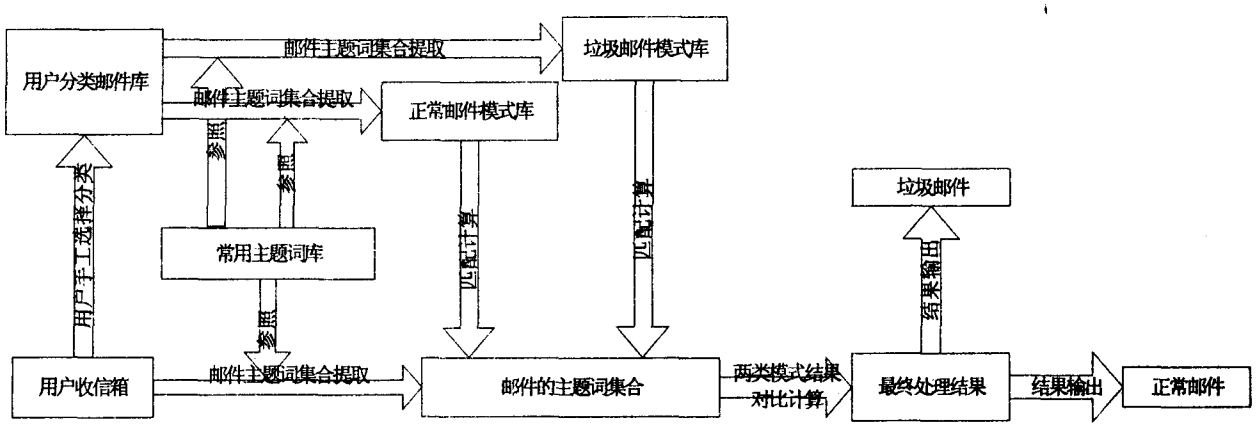

图 1 系统总体的工作框架图

guistic Data consortium(LDC)[9], Oxford Text Archive (OTA)[10]等,根据系统对邮件文本信息处理的具体需要对它们进行相应的一些加工和整理,最终建造出符合系统需要的常用主题词库。

3.3 邮件模式库的建造

这一目标主要通过对邮件用户将最初收到的邮件手工处理的分类结果进行分析来实现。在这一过程中对应于用户手工分拣产生的正常邮件类与垃圾邮件类,分别建造正常邮件模式库和垃圾邮件模式库。在建造过程中使用了类似于生物信息学中对基因片断进行检测的技术,此技术即利用基因信息库中的各种基因信息与待检测基因片断信息进行对照匹配检查来确定待检测基因片断中所有的基因信息。只不过在这里已建造好的邮件主题词库就是"基因信息库",而其中的主题词就是"基因信息",待分析处理用户已分好类的邮件文本内容就是"待检测的基因片断信息"。整个具体建造过程如下:

(1)首先扫描待分析处理的邮件文本内容,将它们与邮件主题词库中的各种主题词进行对照匹配,记录下这些主题词的出现频度。

(2)依照在(1)中记录下的主题词的出现频度针对每一封邮件建立一个矢量 $a_i = \frac{1}{s}(a_{i1}, a_{i2}, \cdots, a_{iN})$,其中的元素为对应主题词在邮件中的出现频度,$s = a_{i1} + a_{i2} + \cdots + a_{iN}$。

(3)将对每一封邮件分析处理生成的矢量按邮件的时间顺序按行排列,最终生成一个矩阵 $\boldsymbol{M}$,$\boldsymbol{M} = \frac{1}{t}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_t \end{pmatrix}$,$t$ 为总的邮件数量。

对用户已分拣好的正常邮件类和垃圾邮件类中的邮件文本按照上面的步骤分别进行分析处理,由此最终生成的矩阵 $\boldsymbol{M}_{\text{normal}}$ 和 $\boldsymbol{M}_{\text{spam}}$ 分别形成正常邮件模式库与垃圾邮件模式库。

在这个数学模型[2]中,笔者认为每一类邮件中一封邮件内容代表一种用户好恶的个性化信息,因而在步骤(2)中生成的矢量即代表一种针对个人的个性信息模式,在这里分别对应着个人的正常邮件模式和垃圾邮件模式。在上面步骤(2)中,对 s 的运算主要为了消除邮件内容长度对单个模式信息的影响,而步骤(3)中对 t 的运算则为了消除一类邮件的数量对此类邮件模式信息的影响。

3.4 邮件的检测识别处理过程

(1)对于用户收到的一封邮件,系统首先利用与上面3.3(1)中从邮件中提取主题词一样的技术从它的文本中提取出主题词出现的频度,建立一个矢量 $\boldsymbol{U} = \frac{1}{s}(u_1, u_2, u_3, \cdots, u_N)$,其中的元素为对应主题词在邮件中的出现频度,$s = u_1 + u_2 + u_3 + \cdots + u_N$。

(2)利用已生成的正常邮件模式库 $\boldsymbol{M}_{\text{normal}}$ 和垃圾邮件模式库 $\boldsymbol{M}_{\text{spam}}$ 分别进行如下计算:

$$T_{1\times N} = \boldsymbol{M}_{\text{normal}}\boldsymbol{U}^{\mathrm{T}}, F_{1\times N} = \boldsymbol{M}_{\text{spam}}\boldsymbol{U}^{\mathrm{T}} \{T_{1\times N} = (t_1, t_2, \cdots, t_N),\ F_{1\times N} = (f_1, f_2, \cdots, f_N)\},\ t = \sum_{i=1}^{N} t_i,\ f = \sum_{i=1}^{N} f_i$$

以上主要考虑了单封邮件模式信息与检测邮件的累计匹配效果。另外有时邮件模式信息之间的组合情况也可能潜在地代表用户的一种邮件模式信息,例如用户在一封邮件中谈到的事情可能涉及以前几封邮件中的内容,对它们的计算如下:

$$T'_{1\times 1} = P.\boldsymbol{M}_{\text{normal}}.\boldsymbol{U}^{\mathrm{T}},\ F'_{1\times 1} = P.\boldsymbol{M}_{\text{spam}}.\boldsymbol{U}^{\mathrm{T}} \{P_{1\times N} = (1,1,1\cdots 1), T'_{1\times 1} = (t'_1), F'_{1\times 1} = (f'_1)\},\ t' = t'_1,\ f' = f'_1$$

不过邮件模式库中的单封邮件模式信息交叉组合产生潜在的邮件模式信息对邮件检测识别的影响要弱于单封邮件模式信息,考虑到这种情况,可以分别对它们赋予不同的影响因子,其中前者小于后者,它们的值一般通过启发式的方法获得。

(3)通过对正常邮件模式库和垃圾邮件模式库的匹配累积结果进行对比计算,识别确定检测的邮件是正常邮件还是垃圾邮件,判别计算如下:

$$G = (\alpha.t + \beta.t')/(\gamma.f + \delta.f')$$

其中 $0 \leqslant \alpha, \beta, \gamma, \delta \leqslant 1, \alpha + \beta = 1, \gamma + \delta = 1, \alpha > \beta, \gamma > \delta$。

若 $G \geqslant \rho$,则检测的邮件为正常邮件,若 $G < \rho$,则检测的邮件为垃圾邮件。在此处的 ρ 为判别邮件是否是垃圾邮件的阈值,它的值可以通过以下训练学习的方法得到。在邮件模式库建好后,选择一些已分好类的邮件样本(正常邮件类和垃圾邮件类),按照上面一样的操作步骤进行计算,将对应正常邮件类样本的邮件产生的 G 值的取值区间与对应垃圾邮件类样本的邮件产生的 G 值的取值区间作对照对比,确定出它们之间的分界线,最终由此分界线得出上面计算所用到的阈值 ρ。

3.5 对两类模式库的更新和整理及其相关维护工作

用户对邮件信息内容的兴趣喜好随着时间可能发生变化,例如原来不需要的邮件,现在可能需要它们了,而原来可以正常接收的邮件,现在不想再接收它们

了,对个人而言,相应地正常邮件类和垃圾邮件类之间产生了一些相互转换。针对这种情况,邮件模式库需要做一些更新调整以反映用户这种邮件个性信息取向的变化。

邮件模式库中的各个邮件模式按时间顺序存储,并且它们的数量维持在一个定量水平,通过向邮件模式库中添加一些用户新选择的分类邮件产生的邮件模式(它们反映了用户新的邮件个性信息),同时系统自动删除邮件模式库中相同数量的最旧的一些邮件模式(它们反映了用户一些过时的邮件个性信息),通过这些操作,系统能够保证邮件模式库始终反映用户当前的邮件个性信息,同时也实现了过滤邮件所用到的用户个性化邮件信息模式的动态更新。由于用户可能对某一类信息特别关注,在一定时期内经常收到同一类信息的邮件,相应地在选择分类邮件建造模式库时添加了这同一类的邮件信息,因此在邮件模式库中生成了多条类似的邮件模式,根据上面的邮件识别过滤原理,这些模式将增强对以后接收到的同类邮件信息(正常邮件或垃圾邮件)的识别准确率,同时也反映了用户当前个性化邮件信息的侧重面,因而邮件模式库中这种类似"冗余"的邮件模式信息不但对整个系统无害,反而能很好地提高整个系统性能。

另外需要提到的是当邮件模式库发生更新变动时,在上面 3.4(3)中计算用到的阈值 ρ 也相应地针对新情况自动重新训练计算产生新值,这样能够实现随用户个性化邮件信息的变化系统能够始终准确地分辨正常邮件与垃圾邮件。

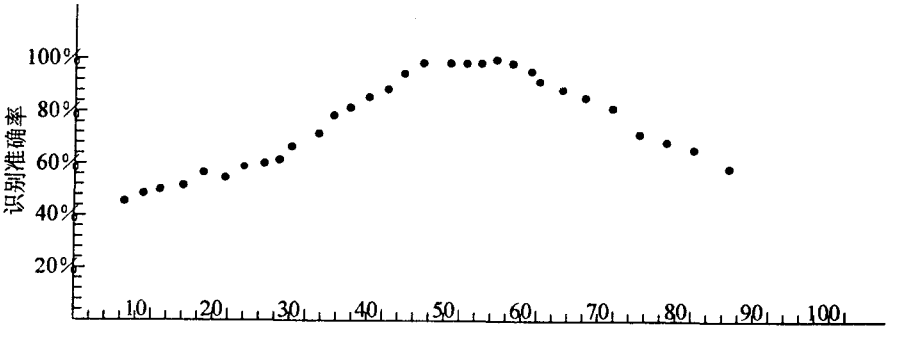

图 2 实验结果统计图

4 系统模型的实现及其相应实验结果分析

通过参照目前常用的一些 E - mail 标准,例如 STMP,POP3 等,利用 Java 中的一个 E - mail 专用 API,即 Java Mail API,同时借助于 MySQL 平台,初步建立了上面这个系统模型的一个原型系统。从功能上讲,此原型系统实际上是一个具有垃圾邮件过滤功能的 MTA(mail transfer agent,邮件传输代理)。

通过对这一原型系统进行一些实际测试和实验,证明它对用户个性化邮件环境中的垃圾邮件具有很好的识别过滤效果,下面是一些主要的实验结果和相应的分析,实验数据是笔者近期的一些日常来往邮件。

在图 2 中,当邮件模式库中的邮件模式数为 31(正常邮件模式数为 11,垃圾邮件模式数为 20),系统的垃圾邮件过滤效果比较差。当邮件模式库中的邮件模式数为 53(正常邮件模式数为 20,垃圾邮件模式数为 33)时,可以看到此系统的垃圾邮件过滤效果比较好,甚至比贝叶斯过滤器还要好一些[1]。当邮件模式库中的邮件模式数增加到 82(正常邮件模式数为 31,垃圾邮件模式数为 51)时,此系统的垃圾邮件的过滤效果反而变差。从图中此系统的整个过滤效果曲线看,此系统针对笔者当前的日常邮件的垃圾邮件过滤效果在邮件模式数处在 47～58 区间时最好,低于这个区间和高于这个区间效果变差,这说明了笔者日常邮件的个性化模式通过当前 47～58 封邮件就可以恰当地反映出来,过少模式信息不足以有效过滤垃圾邮件,而过多模式信息混淆了正常的个性化模式也使得过滤效果变差。以上实验针对的是笔者的日常邮件数据,由于此系统是针对用户个性化邮件信息而设计的,因而每个使用此系统的人获得过滤效果最佳的邮件模式数区间会有所不同,这一点在此系统针对其他一些人的日常邮件信息的实验中得到了证实。因此每个用户在使用此系统时要针对自己的个性化邮件信息来调整获得适合自身的一个具体最佳邮件模式数数值。

5 结束语

针对垃圾邮件具有因人而异的个性化特征,提出了一种基于个性化邮件模式过滤垃圾邮件的系统,经过实验证明此系统针对个人具有很好的垃圾邮件过滤效果,它不同于目前的一些通用垃圾邮件过滤方法,它能更好地适应个人过滤垃圾邮件的个性化需要。

在以后的工作中需要完善和改进系统中用到的具体算法,例如邮件模式匹配的算法等,同时利用一些更好的机器学习的方法更有效地训练系统自动获得系统需要用到的一些量值,另外给系统增添能够识别过滤更多语言邮件文本的功能,从而使整个系统更趋于完美。

(下转第 165 页)

3 结束语

描述了引入时间约束的角色访问控制,更好地满足了访问控制最小权限的原则 ,可以有效地解决时间敏感活动的访问控制问题,增强了访问控制的力度。引入时间后的系统有着更全面、更具体的安全属性描述能力。但是仍然存在许多值得研究的问题,关于时间约束在角色访问控制中的实际应用应进一步进行研究,文中在时间约束的角色访问控制中的整体方面还有待进一步完善。

参考文献:

[1] Sandhu R S, Coyne E J, Feinstein H L, et al. Role—Based Access Control Models[J]. IEEE Computer, 1996, 29(2): 38 - 47.

[2] 张新华,陈军冰. 时间约束的 RBAC 模型及应用[J]. 计算机技术与发展,2007,17(6):246 - 249.

[3] Bertino E, Bonatti P, Ferrar E. TRBAC: A temporal Role - based Access Control Model[J]. ACM Transactions on Informarion and Systems Security, 2001, 4(3): 191 - 233.

[4] Ferraiolo D F, Sandhu R, Gavrila S. Proposed NIST standard for role - based access control[J]. ACM Transactions on Information and System Security, 2001, 4(3): 234 - 274.

[5] 黄　建,卿斯汉,温红子. 带时间特性的角色访问控制[J]. 软件学报,2003,14(11):1944 - 1949.

[6] 胡程瑜, 李大兴. 带时间约束和角色控制的工作流系统授权模型[J]. 山东大学学报,2006,36(3):39 - 43.

[7] 杨　珍,刘连忠. 时间约束的角色访问控制系统的设计与实现[J]. 计算机应用研究,2008,25(1):195 - 199.

[8] 张少敏,王宝义,周利华. 一种具有时间约束的基于角色的授权管理模型[J]. 武汉大学学报,2006,52(5):578 - 581.

(上接第 158 页)

参考文献:

[1] Sahami M, Dumais S, Heckerman D, et al. A Bayesian Approach to Filtering Junk E - Mail. In Learning for Text Categorization[C] // 1998 Workshop. Madison, Wisconsin: [s. n.], 1998.

[2] Duda R O, Hart P E, Stork D G. Pattern Classification 2nd [M]. [s. l.]: Wiley, 2002.

[3] 杨　斌, 路　游. 基于统计学习理论的支持向量机的分类方法[J]. 计算机技术与发展, 2006,16(11):56 - 58.

[4] 张　丽, 黄　东. 基于 Winnow 算法的反垃圾邮件引擎的设计与实现[J]. 计算机技术与发展, 2006,16(4):170 - 175.

[5] 成宝国, 冯宏伟. 一个基于 Naive Bayesian 垃圾邮件过滤器的改进[J]. 计算机技术与发展, 2006,16(2):98 - 99.

[6] 戴劲松,白英彩. 基于贝叶斯理论的垃圾邮件过滤技术[J]. 计算机应用与软件, 2006(1):110 - 111.

[7] 汤　伟, 程家兴, 纪　霞. 一种基于概率推理的邮件过滤系统的研究与设计[J]. 计算机技术与发展, 2008,18(8): 76 - 79.

[8] 龚　伟, 李柳柏. 基于 IDSS 的中文垃圾邮件过滤模型设计[J]. 计算机技术与发展, 2007,17(3):163 - 165.

[9] Linguistic Data Consortium(LDC)[EB/OL]. 2007. http://www.ldc.upenn.edu.

[10] Oxford Text Archive(OTA)[EB/OL]. 2007. http://ota.ahds.ac.uk/.

(上接第 161 页)

需要综合其他多渠道的信息并抽象其特征。另外可以通过对 Snort 规则库的分析,按照协议包头、协议类型、端口等划分、归纳出更多的网络异常特征,然后可以选择适当的串匹配算法,或者特征匹配算法来快速达到定位。

3 结束语

特征检测与异常检测是不能分割开来的,如何能更加有效地通过异常检测的手段来发现问题,结合特征检测的方式来提取异常特征,从而更加快速、准确地查找出异常网络的根源,是我们最终要实现的目标。文中主要是将两种检测手段的优势提取出来,并且把各自的特点相结合,在面对更加复杂的网络异常问题时,可以灵活地作出选择与判断,从而为应用各种模型来实现入侵检测作了比较充足的工作。

参考文献:

[1] 肖海军,王小非,洪　帆,等. 基于特征选择和支持向量机的异常检测[J]. 华中科技大学学报,2008,36(3):99 - 102.

[2] David F,王建新,王　斌. 基于异常和特征的入侵检测系统模型[J]. 计算技术与自动化,2004,23(3):19 - 22.

[3] Sielken R S. Application Intrusion Detection[R]. Virginia: University of Virginia, 1999.

[4] 王平辉,郑庆华,牛国林,等. 基于流量统计特征的端口扫描检测算法[J]. 通信学报,2007,28(12):14 - 18.

[5] 曹苏来. 典型攻击行为描述及特征向量提取[J]. 科技信息,2008(2):37 - 39.

[6] 李韦韦,陈　海,徐振朋. 基于多层特征匹配的网络入侵检测系统[J]. 计算机应用与软件,2008,25(2):278 - 280.

[7] Guimaraes M, Murray M. Overview of intrusion detection and intrusion prevention[C] // Proceedings of the 5th annual conference on Information security curriculum development. [s. l.]: [s. n.], 2008.