基于感知学习的垃圾邮件过滤算法

周 潇,帅建梅
ZHOU Xiao,SHUAI Jian-mei

中国科学技术大学 自动化系,合肥 230027
Department of Automation,University of Science and Technology of China,Heifei 230027,China
E-mail:zhx71@mail.ustc.edu.cn

ZHOU Xiao,SHUAI Jian-mei.Spam filtering method based on cognition learning.Computer Engineering and Applications, 2007,43(28):118-121.

Abstract: The cognition learning model was proposed by Edelman et al.,according to his Theory of Neuronal Group Selection (TNGS) proposed by Edelman et al.presented a cognition learning model.In this paper,we consider the spam filtering problem by analogy to the learning process.In this model,a novel self-learning spam filtering method based on cognition learning is presented.The method uses an improved textual signature algorithm to calculate the similarity matrix of the email text.This matrix,together with the characteristics of email traffic,is regarded as the parameters of the spam detecting method.Finally,the simulation results are presented.

Key words: spam filtering;cognition learning;textual similarity

摘 要:Edelman 等人根据其神经元群选择学说(the Theory of Neuronal Group Selection,TNGS)提出了脑感知学习的模型,将该模型中脑对陌生事物的学习类比于垃圾邮件过滤系统中对未知邮件的学习,提出了一种新的基于感知学习的网络垃圾邮件过滤算法,并将其应用于一种基于合作式网络的垃圾邮件过滤系统模型中。系统使用改进的文本数字签名技术得到邮件文本之间的内容相似度矩阵,将其与邮件到达的行为特征等一起作为该算法的参数,最后给出了仿真实验结果。

关键词:垃圾邮件过滤;感知学习;内容相似度
文章编号:1002-8331(2007)28-0118-04 **文献标识码**:A **中图分类号**:TP393.08

1 引言

据联合国组织机构国际电信联盟统计,全世界目前 80%的电子邮件是垃圾邮件。垃圾邮件一般具有群发的特征,因为群发可以降低垃圾邮件制造者发送邮件的成本,从而给其带来比较高的收益。通过僵尸网络或群发软件产生的垃圾邮件,可能会在发送的过程中对邮件增加无用符号,如空格或标记,这就需要对邮件进行相似度的判断,当相似度超过一定阈值,就可认为是相似的邮件。如果同一封邮件改动的频率很高,则垃圾邮件发送者的代价也就大大增加,很少会出现这种情况。已有很多技术可以在单个反垃圾邮件代理上对垃圾邮件从内容上进行自动过滤,如基于规则的 Ripper 算法、决策树 C4.5 算法、Boosting 方法,基于统计的支持向量机等,其中贝叶斯分类方法[1]使用最广。以上过滤技术缺点主要有:(1)该类技术主要用于邮件用户代理 MUA(Mail User Agent)和因特网邮件传送代理 MTA(Mail Transfer Agent)上,客户端可以设置简单的过滤规则。但是由于 MTA 流量很大,基于内容的过滤会大大影响服务器的效率。分布在各个 MUA 上的过滤系统各自工作,没有利用各个 MUA 可能收到相同内容垃圾邮件这一重要特性。(2)垃圾邮件文本可能被插入一些无用符号,或段落顺序被改变,从而出现很多内容近似的垃圾邮件,导致现有内容过滤技术的正确率大大降低。

合作式垃圾邮件过滤技术可以克服各自工作的内容过滤系统的缺点。该技术主要通过对邮件进行签名来防止垃圾邮件,通过分布在各地的反垃圾邮件代理对垃圾邮件进行实时的判断和签名,服务器端的自动监测系统会将收到的信息和已知的垃圾邮件作比较,这些已知的垃圾邮件是由自动检测系统或最终的收件人检测到的[2],如已经投入商业使用的合作式垃圾邮件过滤系统 Distributed Checksum Clearinghouse[3]。但是在现有的合作式垃圾邮件过滤系统中,每封邮件都要通过服务商服务器上维护的垃圾邮件数据库来判断是否是垃圾邮件,这样消耗的时间和资源是很大的。而且需要用户向服务器报告垃圾邮件,增加了用户的负担,出现恶意用户时,系统的可靠性下降。

本文提出一种新的基于感知学习的垃圾邮件过滤算法,其应用于一种基于合作式网络的垃圾邮件过滤系统模型中。该模型将垃圾邮件的单点内容过滤技术和合作式过滤技术结合,将内容过滤技术应用到组成合作式网络的各个反垃圾邮件代理上,进行第一层过滤,并根据神经元群选择学说中的感知学习理论提出了一种基于感知学习的垃圾邮件判断方法,将其应用

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z449)。
作者简介:周潇(1983-),女,硕士研究生,主要研究领域为模式识别,人工智能;帅建梅(1961-),女,高级工程师,主要研究领域为计算机网络、信息。

在合作式网络的集中处理器上，进行第二层过滤。该基于感知学习的垃圾邮件过滤系统模型可以充分利用网络内的邮件内容重复或相似的信息，并使垃圾邮件检测过程成为一个自学习过程。

2　基于感知学习的垃圾邮件过滤系统模型中关键技术

基于感知学习的垃圾邮件过滤算法是根据现有神经生物学的理论提出的，其认为记忆学习与神经细胞之间的突触生长与变化有关。突触是神经细胞之间传递信息的部分，如果同样的刺激重复多次，神经冲动沿同一神经通路传递的次数增多，突触就会生长、增加，其活性增强，使之与相邻的神经细胞联结、沟通。接受同样的刺激次数越多，其联结就越紧密而形成了固定模式，这时也就形成了记忆。相反，长时间对突触没有刺激，就会使其活性下降，造成遗忘。诺贝尔奖获得者 Edelman G M 的神经元群选择学说在现有神经生物学理论的基础上提出了详细的脑信息处理模型，认为神经系统是以选择而非指令的方式处理信息的，感知学习是在重复刺激模式下的选择强化过程，是根据不断的刺激对神经元群的突触做出修正[6]。本系统是以该学说作为原始的理论模型的。本文中将垃圾邮件的新邮件和已有邮件样本的内容相似度看作是刺激，数据库中的每一项内容看成是突触，将每一项中内容的活性参数看作是突触的活性，并将该活性参数值作为系统判断的依据。样本的内容和已有邮件的相似度越大，则对该突触的刺激也越大，相应其活性也就越强，而邮件为垃圾邮件的可能性也就越大，当其活性不断受到相似邮件的刺激而到达一定的阈值时就将其判定为垃圾邮件，当活性衰减到低于某一阈值时即将其遗忘。对于正常邮件，因为没有新的刺激，或者刺激的间隔时间比较长，其活性衰减到遗忘阈值即被删除。

系统对邮件的判断依次有三个步骤：

(1)基于 hash 值的文本向量化：将文档内容分解，由若干组成文档的特征集合表示。这一步是为了方便后面计算相似度，主要通过 hash 编码等文本向数字串映射方式以方便后续的特征存储以及特征比较，起到减少存储空间，加快比较速度的作用。本系统采用的是 TTTD(Two Thresholds，Two Divisors Algorithm)方法对文本进行分段，然后用非定长子段的 SHA-1 编码对文本进行向量化，生成数字签名序列；

(2)数字签名序列的相似度计算：根据文档特征重合比例来确定是否重复文档。本系统根据生成的文本 hash 值向量，建立文本-分段双向图，然后待分类样本对每个已有垃圾邮件样本建立相似度向量矩阵；

(3)基于感知学习的垃圾邮件过滤算法：根据比较一定时间范围内的邮件内容之间的相似度，判定是否为近似邮件，当内容近似重复的邮件达到一定强度则认为是垃圾邮件。

2.1　基于 hash 值的文本向量化

数字签名是一种公开密钥加密技术的应用。其主要方式是报文的发送方从报文文本中生成一个散列值(或报文摘要)。发送方用自己的专用密钥对这个散列值进行加密来形成发送方的数字签名。然后，这个数字签名将作为报文的附件和报文一起发送给报文的接收方。在判断相同文本时，数字签名的优势十分明显，其使用的 hash 函数主要为 SHA-1 编码，MD5 编码和 Rabin 指纹。尽管该算法可以方便地产生 hash 值并进行比较，但是对于垃圾邮件而言，有一个致命的缺点，即不能处理内容相似的垃圾邮件。本系统使用的是改进的数字签名技术，其基于非定长的文本分段算法，应用 hash 值的数字签名技术，来将邮件进行向量化。

当邮件到达反垃圾邮件代理时，首先要将邮件进行分段，对每个子段进行 hash 值计算。将邮件内容分成大小固定的子段可以减少计算量，但是如果内容有稍微变化(比如插入或者删除一个字符或者单词)，其影响会比较大，而使用变长的分段对增加计算量时，内容变化只是造成局部影响。本系统采用的是 TTTD 非定长分段方法对邮件文本进行分段，该方法是 BSW(Basic Sliding Window Algorithm)分段方法的改进[4]。使用这种分段方法的目的是使子段的长度就被限制在一定的范围内，从而大大降低了该分段方法的相似数据丢失率。

当邮件到达各个反垃圾邮件代理时，系统将分好的子段进行 SHA-1 编码，这样也可以保证用户邮件的机密性[5]。SHA-1 编码的速度很快而且很安全，其产生长度为 160 bits 的散列值，而且使用一个安全的冲突消解算法，使得不同的标志串生成相同的 hash 值的概率低于 2^{-160}，因此抗穷举性很好，而且它适用于任意长的序列的编码。不同的子段生成相同的 hash 值是非常小的。子串的 SHA-1 编码过程完成后，将子串编码作为文本特征向量传送到集中处理器进行匹配处理。

2.2　数字签名序列的相似度计算

在基于向量空间模型(VSM)的文本分类过程中，文本的特征向量与各类代表向量的相似度是判断文本类别的重要标志。本系统中每个样本存储的方式为直接存储，具体如图 1 所示。当每一个邮件样本到达时，系统将其与已有记忆系统中的每一项进行比较，更新每一项的活性。集中处理器的垃圾邮件数据库存储的方式是文本-子段双向分类图[7]。建立该数据结构的步骤：

110	17 fc 2b 87 fe 73 fb 79 ea 5a 40 de de 6a 4a 05 6b 7c ed 75	Email 236
164	2f de a7 a1 a6 9f fe b2 6d f1 c2 67 c2 cd f2 53 86 9b cd ab	Email 236
144	82 98 38 3f de 6b ac 44 a0 47 96 f4 2b 3e c1 10 ea a7 31 b7	Email 236
95	99 45 45 e0 09 45 75 c9 a7 23 30 32 e2 9f 68 4b 9f f0 c8 e3	Email 236
139	54 f0 17 ff 33 4e 32 8d 5c 14 35 64 ec a1 21 4e f6 bd 14 8f	Email 236
166	d2 71 30 38 eb 3e 4c 31 9d b2 2d 69 8e 2c 0d 3a ad f8 4f 8c	Email 236

图 1　子段 hash 编码

(1)将每一个子段用<子段长度，hash 值，邮件编号>的形式表示。形式如图 1 所示。

(2)如果同时需要处理大量的数据，可以应用硬盘分类的技术，如并行计算，使得出现相同子段的文件出现在数据库中相临的位置。

(3)用 union-find 聚类方法以 hash 值为特征对各项进行聚类，将出现相同 hash 值的文件表示在一行中。

当一个需要判断的邮件特征向量到达时，即对每一维特征对存储树进行搜索，S_j 表示新邮件和已有第 j 个样本邮件的相似性，N 表示新邮件的子段数，两个邮件出现相同的子段就将，S_{j+1}，全部搜索结束后，$S_{j/N}$ 即得到了新邮件和已有第 j 个样本邮件的相似度。从而建立起新邮件对已有样本项的相似度矩阵。系统通过该相似度进行计算。M 表示样本邮件数据库中子段的数目，则建立该存储树的时间复杂度为 $O(M\log M)$，查找一个子段的复杂度为 $O(M\log M)$，因为邮件中子段的个数是线性的，所以一封邮件在系统中对每个样本判断相似度的时间复杂度为 $O(M\log M)$，当存储的邮件样本增加时，不会增长的很快。

2.3 基于感知学习的垃圾邮件过滤算法

根据 TNGS 对突触活性的修正公式[6],本系统的垃圾邮件样本活性变化函数表示成:

$$V_i(N)=\omega\cdot V_i(N-1)+(S(u)-\theta)\cdot\Phi(\nabla t_N) \quad (1)$$

$V_i(N)$表示项内有 N 封邮件时的活性,ω 表示衰减系数,这里用 $\omega=e^{-\frac{\Delta t_N}{\tau}}$,$\Delta t_N$表示两次邮件到达的时间间隔,$\tau$ 表示衰减时间常数,$S(u)$表示输入邮件 u 和项内样本的相似度,θ 表示认为新邮件与样本为相似邮件的相似度阈值,当相似度超过该阈值时活性才更新。$\Phi(\nabla t_N)$表示以邮件间隔时间 ∇t_N为参数的函数,∇t_N很小时,且持续时间比较短时,本系统认为出现正常群发邮件的可能性比较的大。本模型选择 $\Phi(\nabla t_N)=(\nabla t_N)^\gamma$,$(0<\gamma<1)$。$0<\gamma<1$ 时,$(\nabla t_N)^\gamma$是一个凹函数,表示新到相似邮件对样本活性的刺激强度随着邮件到达时间间隔的增加而增加,而当发送间隔比较大时,该项的增量减慢,这样就可以让发送周期比较长而又可能相似的正常邮件顺利通过。V_{spam}为记忆阈值,当项活性超过该值时,该项判定为垃圾邮件。

大量相似的垃圾邮件往往不是同一时间到达的,而是持续几小时甚至几天,所以本系统引入了时间参数,这样就使判定系统具有了自学习记忆功能。hash 表中每一项都有一个域是表示该项活性的,当一个相似的邮件序列到达时,通过计算相似度,新的活性则通过增强函数计算得到。每过一定的时间数据库就利用衰减函数进行更新,当表项的活性增加到一个阈值时即判定其为垃圾邮件,当邮件的活性低于某一阈值时就将该邮件样本删除。

3 基于感知学习的垃圾邮件过滤系统结构

垃圾邮件过滤系统的结构如图 2 所示,共分两层,第一层过滤在小型合作式网络中的各反垃圾邮件代理端,第二层在集中处理器中。

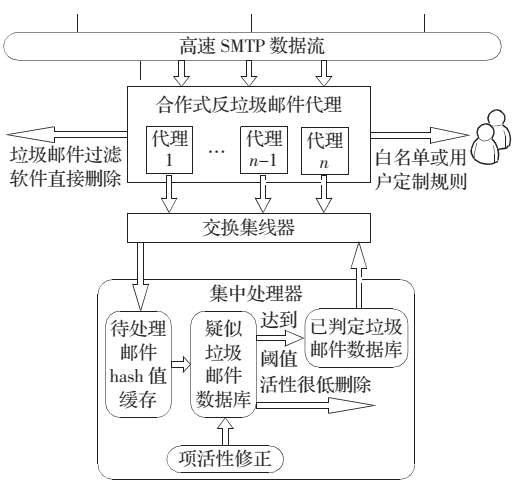

图 2　系统结构图

第一层过滤在组成合作式网络的反垃圾邮件代理端进行。SMTP 处理器首先定位邮件,并重组邮件内容。每一个反垃圾邮件代理都有自己的垃圾邮件检测软件,并建有自己的垃圾邮件特征库,一般的垃圾邮件可以通过黑白名单、贝叶斯内容识别等技术过滤,用户还可以根据自己的爱好定制过滤规则,现有的垃圾邮件过滤软件可以达到要求。反垃圾邮件代理过滤的邮件,一部分根据已经有的过滤规则被直接过滤掉,一部分根据用户的白名单或内容定制直接提交用户,另一部分无法准确判断的被归为可疑邮件,送到集中邮件处理器。为了保护用户邮件的隐私,反垃圾邮件代理用 TTTF 方法对邮件进行分段,然后用 SHA-1 将子段进行编码,由此生成的 hash 指纹序列发送到集中处理器进行处理。

第二层过滤在集中处理器中进行。需判断的邮件样本被送到可疑垃圾邮件数据库中,和已有的样本进行比较,更新各个样本的活性值。如果没有在相似度阈值以上的已有样本,就存储新样本,当已有样本有超过相似度阈值的,就将该样本的分析特征通知各个反垃圾邮件代理,更新各个反垃圾邮件代理的垃圾邮件特征库,并在可疑垃圾邮件数据库中删除该样本。当可疑垃圾邮件数据库中邮件样本的活性低于一定阈值时,直接将其删除。

4 仿真实验结果

本文模拟了一个系统中的垃圾邮件集中处理器,输入的邮件内容相似度数据可以利用 Forman 等人给出的文本相似度技术[7]得到,实验中主要根据现有的邮件传输行为模型调整了该项活性修正函数,当超过记忆阈值 V_{spam} 时,则判定为垃圾邮件。本文中提到的判别代价指的是检出垃圾邮件之前无法判别的邮件数量。根据 Gomes 等人的观察,垃圾邮件的达到数量是服从参数改变的 Poisson 分布的[8]。在实际中,考虑到群发邮件的发送频率很大且具有一次性,正常邮件的发送周期应远远大于垃圾邮件,因此主要考虑过滤发送间隔时间的期望值在几秒到几小时之间的大量群发邮件。当邮件发送持续时间不长,但发送的时间间隔比较短,通过调节 $\Phi(\nabla t_N)$,即使是大量发送的正常邮件也可以通过。对于发送周期比较长的邮件,比如工作报告等,可以根据参数来调节衰减时间,只要发送时间间隔比较长,如几天一封,也可以正常通过。当某样本的相似邮件数目较多时,其为垃圾邮件的可能性也比较大,其活性增长速度增加,通过 $\Phi(\nabla t_N)$中的 γ 参数可以对此调节。

在模拟系统中,垃圾邮件的到达是一个 Poisson 过程,从而邮件到达时间的间隔是服从指数分布的,其概率密度函数为 $P(x)=\lambda e^{-\lambda x}$[9],这里 λ 表示一小时中收到邮件数量的期望值。图 3 考察不同的 λ 下识别一封垃圾邮件的代价,给出了 100 次仿真的平均结果。这里取式(1)中的活性衰减时间常数 τ=2 天,设置同一封垃圾邮件之间相似度为 100%。可以看到,当邮件发送频率小于每天三封时,其判为垃圾邮件的代价很大,基本可以正常通过。发送频率在每小时一封以上,其判别代价随着频

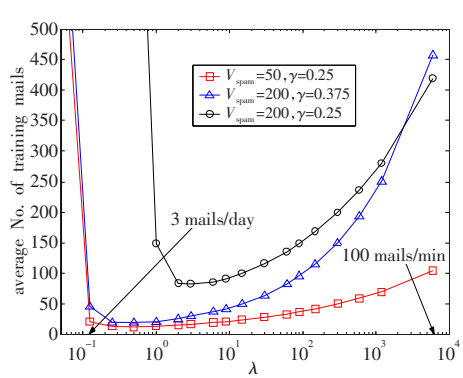

图 3　不同发送频率邮件的判别代价

率的增加而增加，但判别时间却是减小的，每小时一封时的判别时间为 13 h~21 h，而每 12 s 一封的判别时间为 6 min~29 min。V_{spam} 比较高时，判别的代价也增高。γ 主要控制发送频率比较大和比较小时候的判别代价。从图中可以看到，在 V_{spam}=50、γ=0.25 以及 V_{spam}=200、γ=0.375 两种参数设置下不同发送频率邮件的判别代价在实际中都是可行的。用户也可以根据需要灵活的调节这两个参数。

当同一封垃圾邮件因为在网络中的转发而内容稍有变化时，副本之间的相似度有所下降，当下降到低于一定的阈值时则认为是一封新邮件，另行存储，考虑到垃圾邮件制造者的发送成本，邮件内容在一定时间之内的改变不会太大，模拟系统中随机生成了副本之间的相似度。在图 4 中，给出了参数为 V_{spam}=50、γ=0.25 下相似邮件和相同邮件的判别代价。此时判断一封垃圾邮件的代价增大了，不过在实际中仿真的结果仍是可行的。

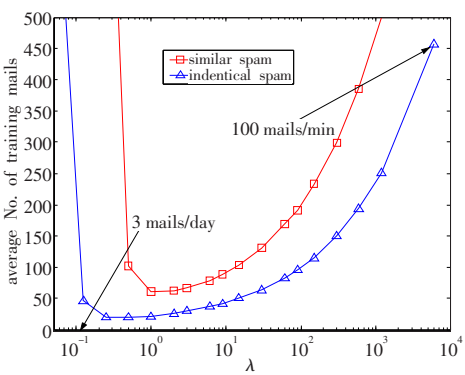

图 4　相同邮件和相似邮件的判别代价

Gomes 等人还指出，对于正常邮件，白天邮件的发送频率应远远大于晚上，而垃圾邮件的发送频率在一天中分布比较平均，在模拟系统中参考使用了在文献[8]中给出的邮件流量数据。仿真实验中模拟了三个用户，用户 1 发送内容完全不同的正常邮件，用户 2 一次性发送群发邮件 100 封，用户 3 发送每天内容改变的垃圾邮件，其邮件的相似度以每天 80%的速度衰减。使用本文中的方法计算相似度，用 TTTF 方法设定的分段有 20%被改动时，其与原文的相似度为 80%左右，相当于实际中经常出现的邮件内容增加(减少)几个字符或某个段落被替换。表 1 给出了系统参数 V_{spam}=200、γ=0.375 时，模拟运行 10 天后的检出情况，结果证明本系统可以有效的检测出每天内容变化的近似垃圾邮件，而正常的群发邮件可以正常通过。

表 1　系统运行情况

	发送频率/s	λ	正确率
用户 1	0.1~0.3	360~1 080	100%
用户 2	100/min	–	100%
用户 3	0.22	800	96.4%

在实际中，系统检出的也可能有正常发送量比较大的邮件，比如用户定制的一些邮件杂志或信息等，可以通过白名单的方式在反垃圾邮件代理端设置让其通过，是容易实现的。本文没有讨论不同用户对垃圾邮件的不同定义情况。

5　结语

本文提出了一种基于感知学习的反垃圾邮件算法，将其应用在合作式网络中。该算法在现有垃圾邮件内容过滤技术的基础上，增加了基于文本数字签名技术的近似内容邮件文本相似度计算，以及对邮件到达频率的判断，通过基于感知学习的邮件样本活性修正公式可以有效的检测出内容相似的垃圾邮件，对发送每天内容改变的垃圾邮件过滤正确率达到 96.4%。其优点首先是不仅可以检测内容重复的垃圾邮件，对于发送过程中内容有所改变的同样适用。其次是在检测时不影响发送周期较长和短时间内群发的正常邮件。第三，集中处理器报告内容大量重复的邮件，将该类邮件生成唯一识别的摘要，通知个单点反垃圾邮件代理更新数据库，大大提高了邮件检测的时间，并保证了邮件内容的保密性。第四，本系统中用户的参与不是必须的，当用户收到垃圾邮件时可以在各自的反垃圾邮件代理上增加黑名单或过滤规则，没有必要向集中处理器或者其他代理汇报，这样就可以避免用户因为判定的垃圾邮件标准不同而造成的误判，而且可以减少恶意用户的虚假报告，增加了垃圾邮件判定的客观性，应用在实际中可以大大扩展有垃圾邮件过滤技术的适用范围。(收稿日期：2007 年 5 月)

参考文献：

[1] Sahami M，Dumais S，Heckerman D，et al.A bayesian approach to filtering junk E-mail[C]//Proceedings of the AAAI Workshop on Learning for Text Categorization，1998：55-62.

[2] Balabanovic M，Shoham Y.Fab：content-based，collaborative recommendation[J].Communications of ACM，1997，40(3)：66-72.

[3] Distributed checksum clearinghouse[CP/OL].[2007].http://www.rhyolite.com/anti-spam/dcc/.

[4] Kave Eshghi，Hsiu Khuern Tang.A framework for analyzing and improving content-based chunking algorithms，HPL-2005-30R1[R].Hewlett-Packard Development Company Report，2005.

[5] National Institute of Standards and Technology.FIPS PUB 180-1.Secure Hash Standard[S]//Federal Information Processing Standards Publication 180-1，1995.

[6] Edelman G M.Neural darwinism[M].New York：Basic Books Inc，1987.

[7] Forman G，Eshghi K，Chiocchetti S.Finding similar files in large document repositories[C]//The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'05)，Chicago，IL，USA，2005：394-400.

[8] Gomes L H，Cazita C，Almeida J，et al.Characterizing a Spam Traffic[C]//Proc 4th ACM SIGCOMM Conference on Internet Measurement.Taormina，Italy：ACM Press，2004：356-369.

[9] Garcia A L.Probability and random processes for electrical engineering[M].2nd ed.[S.l.]：Prentice Hall，1993.