·热点与综述·　　文章编号：1000-3428(2023)05-0038-10　　文献标志码：A　　中图分类号：TP309

基于改进PCFG算法的口令猜测方法

李静雯，赵　奎

（四川大学 网络空间安全学院，成都 610065）

摘　要：近年来口令泄露事件频出，有效的口令猜测方法是保障口令安全的重要手段，其中基于概率上下文无关文法(PCFG)的口令猜测方法效果尤为显著，然而仍存在无法生成新的口令字符子段、对生成口令的概率估计不准确等问题。以基于PCFG的口令猜测方法为研究对象，对其在口令构造过程中关键阶段的命中率进行分析，提出基于Backoff-RNN与概率平衡的改进PCFG口令猜测方法。在口令结构划分阶段，通过分析用户在构造口令时的行为与偏好，将口令从汉语拼音和英文单词两方面进行更细粒度的结构划分，提取口令更深层次的结构信息。在口令填充阶段，将Backoff思想应用于字符级RNN模型，生成子结构中长序列字符子段，提高模型准确性和泛化能力。在口令概率计算阶段，改进口令生成概率的计算方法，解决了使用传统计算规则时因口令结构长度不一致造成的概率不平衡问题。实验结果表明：在中英文两种语言环境交叉数据集上，该方法的漫步口令猜测攻击命中率相较于基于PCFG的口令猜测方法分别提升了20.6%和22.4%；在中文语言环境数据集上，定向口令攻击命中率相较于TarGuess-I模型提升了2.8%。

关键词：口令猜测攻击；自然语言处理；概率上下文无关文法；深度学习；口令安全

开放科学（资源服务）标志码（OSID）：

中文引用格式：李静雯，赵奎.基于改进PCFG算法的口令猜测方法[J].计算机工程，2023，49(5)：38-47.

英文引用格式：LI J W，ZHAO K. Password guessing method based on improved PCFG algorithm[J]. Computer Engineering，2023，49(5)：38-47.

Password Guessing Method Based on Improved PCFG Algorithm

LI Jingwen，ZHAO Kui

（School of Cyber Science and Engineering，Sichuan University，Chengdu 610065，China）

【Abstract】Recently，password leaks have been occurring frequently. Accordingly，effective password guessing methods are an important means of securing passwords，among them，the method based on Probabilistic Context-Free Grammar（PCFG）is extremely effective. However，this method still has problems such as the inability to generate new substrings of password and inaccurate estimation of the probability of generating passwords. Thus，taking the PCFG-based password guessing method as the research object，its hit rate in the key stage of the password generation process is analyzed. Subsequently，an improved PCFG password guessing method based on Backoff-Recurrent Neural Network（Backoff-RNN）model and probability balance is proposed. In the password structure division stage，by analyzing the user's behavior and preference when constructing the password，the password is more finely divided into Chinese Pinyin and English words to extract a deeper structure information. In the password filling stage，the idea of Backoff is applied to the char-RNN model to generate long sequence substrings in the substructures to improve the accuracy and generation ability of the model. In the password probability calculation stage，the calculation method of password generation probability is improved to address the probability imbalance problem caused by inconsistent password structure length when using the traditional calculation rules. The experimental results demonstrate that the hit rate of the trawling attack of the proposed method is 20.6% and 22.4% higher than that of traditional method based on PCFG on the cross-datasets of Chinese and English language environments，respectively，and 2.8% higher than that of TarGuess-I model of the targeted attack on the dataset of Chinese language environment.

【Key words】password guessing attack；natural language processing；Probabilistic Context-Free Grammar（PCFG）；deep learning；password security

DOI：10. 19678/j. issn. 1000-3428. 0064678

基金项目：国家自然科学基金（U19A2068，61872254）。

作者简介：李静雯（1998—），女，硕士研究生，主研方向为大数据安全、口令安全；赵　奎（通信作者），教授、博士。

收稿日期：2022-05-11　　**修回日期：**2022-06-20　　**E-mail：**zhaokui@scu.edu.cn

0　概述

口令认证凭借实现简单、成本低、效率高的特点在众多身份认证方式中占据主流地位[1]。近年来，随着口令泄露事件的频发，口令安全问题突出。从攻击者的角度进行口令猜测攻击，对保障用户口令安全具有重要意义，但其难点在于从已泄露的大规模口令样本中挖掘出用户普遍的口令构造方式。口令生成任务可看作文本生成任务，但口令具有结构化明显、语义语法弱的特点[2]。基于概率上下文无关文法（Probabilistic Context-Free Grammar，PCFG）[3]的口令猜测方法利用该特点在对真实口令数据进行统计分析的基础上，对口令结构及各结构子段进行频次统计，以此生成新的口令，命中率较高且应用广泛。但是，目前基于PCFG的口令猜测方法[4-6]仍存在以下3个方面的问题：第一，在口令结构划分阶段，仅从字符类型的角度对口令进行分割，忽略了字符串中更细粒度的信息，且未对中文和英文进行对比分析与提取；第二，在口令填充阶段，无法生成结构中新的口令子段，虽有学者结合Markov模型[7-9]或循环神经网络（Recurrent Neural Network，RNN）[10]来解决这一问题，但仍面临模型训练时序列长度难以确定的问题；第三，在口令概率计算阶段，将一条口令中各子段的概率累积作为其生成概率，所生成的概率受口令结构长度的影响较大，造成概率计算不平衡。

为了更好地抽象口令的基础语法结构并提高算法对口令子段的命中率，本文提出基于Backoff-RNN与概率平衡的PCFG口令猜测方法。首先，在分析4个大规模口令数据集（2个中文和2个英文口令数据集）的基础上，挖掘出中英文语言环境下口令构造的差异，对口令从汉语拼音、英文单词上进行更细粒度的划分，提高模型的准确性。然后，将Backoff[11]的思想引入字符级RNN模型中，在生成长序列口令子段时，根据已生成的子串动态选择适合长度的RNN模型，均衡模型拟合问题，使得生成的子段更符合真实训练样本中的序列关系。最后，将困惑度（Perplexity）计算方法引入口令生成概率的计算规则，使得改进后的概率计算方法更能体现出口令在口令集中的真实分布规律。

1　相关工作

口令猜测方法根据攻击过程中是否利用用户个人信息可以被分为漫步攻击（trawling attack）以及定向攻击（targeted attack）。前者不针对特定的攻击对象，目标是在允许的猜测次数下提高模型对攻击样本的命中率。后者是在给定目标用户个人信息的前提下，以更少的猜测次数（$\leqslant 10^4$），有针对性地猜测该用户的真实口令。针对上述2种攻击方式，目前主流的口令猜测方法可分为基于马尔可夫（Markov）[7]、基于PCFG[3]、基于神经网络[12-15]等3类。

1.1　基于Markov的口令猜测方法

NARAYANAN等[7]于2005年提出一种基于Markov的口令猜测方法，该方法根据字符序列之间的转移概率来指导口令的生成。在n阶Markov模型中，下一个字符出现的概率基于它前面长度为n的子串。以4阶Markov模型为例：在训练阶段，口令“Li1234”需要统计出首字符“L”的频数，“L”后字符“i”的频数，“Li”后字符“1”的频数，“Li1”后字符“2”的频数，“Li12”后“3”的频数，“i123”后“4”的频数，在遍历训练集中的每个口令后便可得到各子字符串之间的转移概率矩阵；在生成阶段，将各部分概率累乘便可得到目标口令的生成概率，计算公式如式（1）所示；在所有口令生成结束后，将已生成的口令按概率大小进行排序，生成口令攻击字典。

$$p(\mathrm{Li1234})=p(\mathrm{L})\times p(\mathrm{i}|\mathrm{L})\times p(1|\mathrm{Li})\times p(2|\mathrm{Li1})\times p(3|\mathrm{Li12})\times p(4|\mathrm{i123}) \tag{1}$$

基于Markov的口令猜测方法可以反映出字符之间的序列关系，但其只是简单地对已有样本进行概率统计，无法学习字符序列间的高阶特征，模型生成效果易受训练数据影响，产生过拟合，同时该方法也忽略了口令重结构轻语义的特征。

1.2　基于PCFG的口令猜测方法

1.2.1　实现思路

PCFG算法[3]将整个口令段划分为字母子段（L_n）、数字子段（D_n）和特殊字符子段（S_n），其中n表示该子段的长度，如口令“Li#123456”可被划分为基本结构$L_2S_1D_6$及子段“Li”“#”“123456”。

如图1所示，在训练阶段，先提取出训练集中所有口令的基本结构以及被分割出的子段，再在此基础上构建口令结构及各子段的频率字典。

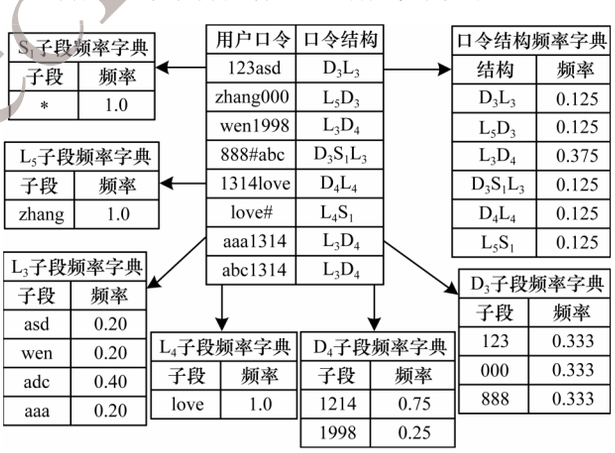

图1　基于PCFG的口令统计过程

Fig.1　Password statistics process based on PCFG

在口令生成阶段，根据上述频率字典，按照频率大小依次取出口令结构对其进行填充。例如，首先从口令结构频率字典中提取出当前频率最高的结构L_3D_4，再分别取出L_3及D_4子段表中当前频率最高的子段“abc”和“1314”进行填充，口令“abc1314”的生成概率计算规则如式（2）所示，以此构建按生成概率

递减的猜测口令集。

$$p(\text{abc1314}) = p(\text{S} = \text{L}_3\text{D}_4) \times p(\text{L}_3 = \text{abc}) \times p(\text{D}_4 = 1314) \quad (2)$$

其中:S表示口令结构。

1.2.2 现有基于改进PCFG的口令猜测方法

VERAS等[16]在字母子段上使用分词、词性标注、归一化等自然语言处理工具分析字母子段的语义特征。HOUSHMAND等[17]在PCFG算法的基础上加入键盘词模式和多字模式以提升命中率。LI等[18]和WANG等[19]结合用户姓名、生日、邮箱等个人信息,分别提出Personal-PCFG和TarGuess-I模型,实现定向猜测用户口令。章梦礼等[5]考虑到PCFG算法无法生成结构中新的子段的弊端,将PCFG算法的结构划分优势和Markov模型在表示前后字符依赖关系中的优势相结合提出SPSR模型,有效地提升了口令猜测方法的准确性和泛化能力。

基于改进PCFG算法的口令猜测方法能够准确地抽象出口令的结构,且生成的口令可按照口令概率排序。但现有的概率计算方法使得口令的生成概率受口令结构长度的影响较大,导致概率计算不平衡。经实验发现,PCFG算法对于字符序列长度在4以内的短序列命中率较高,但随着字符序列长度的增加,命中率会不断降低,因此运用此特征指导口令生成是本文研究的重点。

1.3 基于神经网络的口令猜测方法

近年来,生成式对抗网络(Generative Adversarial Networks,GAN)、RNN在文本生成领域的不断发展为口令生成提供了新思路。HITAJ等[20]提出PassGAN模型,利用GAN来学习真实口令的构造规则,模拟生成猜测口令集。但是,GAN模型无法给出生成口令的概率并且针对文本这类离散数据无法进行反向传播,因此基于GAN的口令猜测方法在命中率上低于传统方法。MELICHER等[10]使用RNN进行口令破解,周环等[21]将个人信息与RNN相结合,提出定向口令猜测模型TPGXNN。另外,长短期记忆(Long Short-Term Memory,LSTM)网络作为RNN的变种,具备长期记忆能力,XU等[22]使用LSTM进行口令破解,LIU等[23]提出PL模型,将PCFG与LSTM结合进行口令猜测。上述基于循环神经网络的口令猜测方法和流程与Markov模型类似,但通过前者得到的概率分布要比基于概率统计的Markov模型更加合理,可提取字符间的隐含特征,因此上述方法相较于传统方法有效地提高了猜测模型的命中率。

LSTM通过引入多种门操作解决了长期依赖问题,但与此同时也增加了模型结构的复杂度及计算量。对于口令数据而言,1条口令的长度通常被限制在8~20,再经过PCFG算法的结构拆分后,子串的长度只会更短,在训练和生成的过程中并不会存在明显的长期依赖问题。汪定等[24]在PL模型[23]的基础上,将其中的LSTM模型替换为RNN提出PR模型,实验结果显示后者的命中率普遍略高于前者,同时训练效率远高于前者。因此,在结合PCFG算法时,RNN模型相较于LSTM在口令生成中更具优势。但由于现阶段无论是Markov、RNN还是LSTM模型,都是在选定模型阶数后进行口令数据的训练与生成,因此均存在概述中所述的弊端。

2 用户口令行为分析

2.1 用户口令数据集

由于口令猜测方法建立在对大规模真实用户口令集的分析工作的基础上,因此选取2个中文背景的用户真实口令数据集和2个英文背景的用户真实口令数据集进行统计分析,综合对比中英文语言环境用户在构造口令时的习惯差异,为汉语拼音和英文单词划分策略提供依据。采用的4个口令数据集的基本信息如表1所示。

表1 口令数据集中的基本信息

Table 1 Basic information in the password datasets

名称	类型	口令数量/条	用户语言	包含个人信息
12306	铁路系统	129 303	中文	是
CSDN	程序员论坛	6 428 632	中文	否
Rockyou	社交网络	32 581 870	英文	否
Yahoo	门户网站	5 626 485	英文	否

2.2 中英文语言环境下常用字符串统计

对数据集中的每一个口令提取长度为3~10的口令子串序列,构造出如表2和表3所示的中英文常用口令子串。由表2和表3可以看出:中文语言环境用户会更频繁且集中地使用简单的字母串和数字串;英文语言环境用户在构造口令时,除了简单数字序列以外,更多地选择常用的英文单词,例如baby、love、one等。

表2 中文语言环境下用户常用口令子串

Table 2 Common password subsegments for users in the Chinese language environment

序列长度	常用子串
3	123,520,111,000,aaa,abc,asd,qwe,666,888
4	aini,1314,1234,love,woai,6666,8888,a123
5	12345,aaaaa,66666,11111,00000,a1234,woshi,ilove
6	999999,000000,321321,112233,qwerty,121212
7	1234567,5201314,zxcvbnm,a123456
8	87654321,aaaaaaaa,12345678,11111111,00000000,888888888,66666666,iloveyou,password,1q2w3e4r
9	123456789,987654321,a12345678,qwertyuiop
10	woaini1314,a123456789,0123456789,1234567890

表3　英文语言环境下用户常用口令子串

Table 3　Common password subsegments for users in the English language environment

序列长度	常用子串
3	123,and,all,000,one,ass,son,aaa,ann,her,111
4	love,1234,ball,baby,1111,ever,rock,life,a123,hell
5	ilove,12345,angel,hello,11111,lover,jesus,lucky
6	monkey,qwerty,prince,dragon,christ,jordan,flower
7	1234567,welcome,michael,diamond,charlie,anthony
8	password,princess,sunshine,iloveyou,november
9	123456789,butterfly,chocolate,Elizabeth,beautiful
10	basketball,tinkerbell,strawberry,volleyball,sweetheart

2.3　中英文语言环境下汉语拼音与英文单词的占比统计

本节对中英文语言环境下所有包含字母的口令中包含汉语拼音、英文单词、混合字母串的口令占比进行统计分析，如表4所示，其中混合字母串表示同时包含汉语拼音子串、英文单词子串、普通字母子串这3种子串中任意2种及以上的一段连续字母子串。由表4可以看出：

1）在12306数据集所有包含字母的口令中包含汉语拼音的口令约占41.5%，在CSDN数据集中这一比率高达73.1%，英文单词在以上两口令集中的占比均达到29.7%和43.6%。在Rockyou和Yahoo数据集中英文单词的口令占比分别达到了49.2%和76.3%。这一统计结果说明汉语拼音在中文语言环境用户的口令构造以及英文单词在英文语言环境用户的口令构造中占据重要的地位。基于上述分析结果，笔者认为对口令中的汉语拼音和英文单词进行提取和标注将会有针对性地提高口令破解的命中率。

2）在4个数据集中，所有包含字母的口令中包含混合字母串的口令占比均在50%左右，在这种情况下基于PCFG的口令猜测方法（简称为传统PCFG方法）仅将字母段提取为以长度为单位分割的字母子段L_n，无法体现出用户口令构造时的深层结构特征，因此本文在此统计结果的基础上，对字母子段从汉语拼音和英文单词两方面进行更细粒度的标记与提取。

表4　口令数据集中汉语拼音和英文单词的占比统计

Table 4　Statistics on the proportion of Chinese Pinyin and English words in the password dataset　%

数据集名称	所有包含字母的口令占比	包含汉语拼音的口令占比	包含英文单词的口令占比	混合字母串的口令占比
12306	72.5	41.5	29.7	47.6
CSDN	56.8	73.1	43.6	49.2
Rockyou	92.3		49.2	49.4
Yahoo	94.3		76.3	47.1

3　基于Backoff-RNN与概率平衡的PCFG口令猜测方法

本文提出的基于改进PCFG算法的口令猜测方法（简称为所提方法）由口令结构划分、多阶RNN模型训练、口令生成等3个模块构成，如图2所示。首先，将数据集中的所有口令按照个人信息、汉语拼音、英文单词、字母子段、数字子段和特殊字符子段进行细粒度分割，得到口令结构及各子段的频率字典，并将各子段作为多阶RNN模型的训练数据。然后，利用Backoff-RNN模型生成长度大于4的字母子段、数字子段、特殊字符子段。最后，依次对口令结构字典中的口令结构进行填充，利用改进的概率计算规则对生成口令进行概率计算，并按照概率从高到低的顺序对口令进行排序，生成口令猜测字典。基于Backoff-RNN与概率平衡的PCFG口令猜测方法在保留了传统PCFG方法优势的基础上对其在口令结构划分、生成长序列子段及概率计算上存在的不足进行改进。

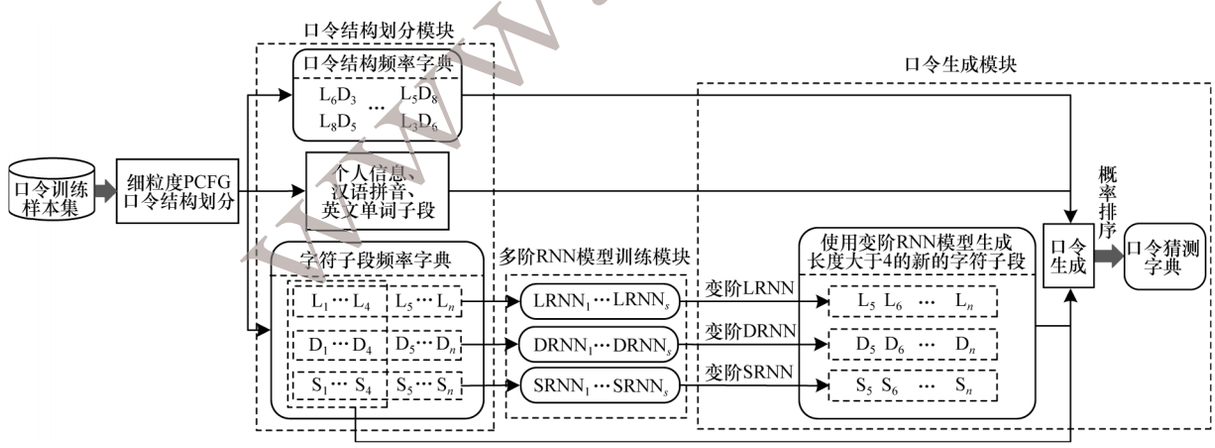

图2　基于Backoff-RNN与概率平衡的PCFG口令猜测方法

Fig.2　PCFG password guessing method based on Backoff-RNN and probability balance

3.1　口令结构划分模块

在传统PCFG方法的基础上进行更细粒度的结构划分，将用户口令依次从PI（个人信息）、P（汉语拼音）、W（英文单词）、L（字母子段）、D（数字子段）、

S(特殊字符子段)等6个类别中进行口令结构提取:

1)个人信息。在进行定向口令猜测攻击时,借鉴TarGuess-I[19]中对用户个人信息的划分策略,从姓名、生日、邮箱前缀中提取口令中的个人信息,但与其划分规则稍有不同的是,本文未在"年月日""月日年"这类仅调换序列顺序的数据上进行抽取,而是通过将其拆分为"年""月日""日月"等更细粒度的字段,以容纳用户更多形式的结构变换。

2)汉语拼音。将一个汉语拼音看作一个整体,在统计时只对汉语拼音的个数进行提取,例如"mima"将被抽象为P_2。

3)英文单词。搜集常用英文单词、英文名、地名等构成英文词典,使用和汉语拼音同样的处理方式对英文单词序列进行提取。

4)将余下的字符串分别替换为字母子段、数字字段及特殊字符子段。

至此,便完成了口令的结构划分,得到了口令结构以及各子段的频率字典。

3.2 多阶RNN模型训练模块

利用上一阶段拆分出的数字子段、字母子段、特殊字符子段分别训练出基于数字的$DRNN_i$、基于字母的$LRNN_i$及基于特殊字符的$SRNN_i$($i\in[1,s]$,s为变阶RNN模型中的最高阶数),并提出基于Backoff思想[11]的自适应变阶RNN模型来生成口令子段。在生成子段时根据已生成子串,自动选择RNN模型的阶数来预测下一位生成字符。

模型训练过程与字符级RNN模型相似,具体过程为:在预处理阶段,对提取出的每个子段尾部添加结束符〈EOS〉;在训练阶段,对于i阶RNN模型RNN_i,选择长度不小于i的子段(不包含结束符)作为其训练数据,对于每一个子段,从第一个字符开始,以滑动窗口的方式,截取窗口大小为i的子串作为模型的输入序列,并将当前窗口后的下一个字符作为标签,滑动窗口以1为步长不断向后滑动,直到获取到的字符标签为结束符。以训练基于数字的$DRNN_i$模型为例,训练过程如图3所示。

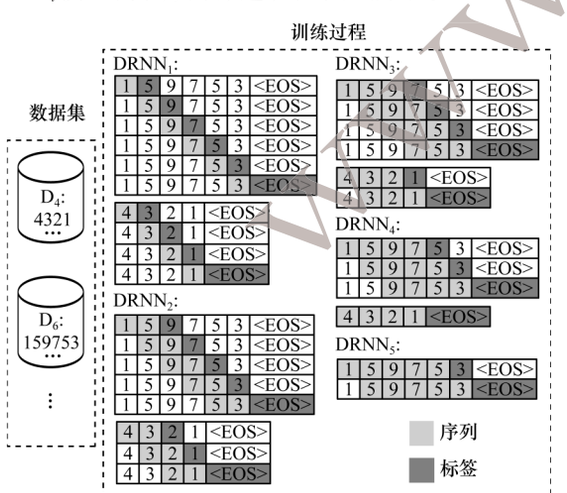

图3 模型训练示例

Fig.3 Example of the model training

3.3 口令生成模块

3.3.1 Backoff-RNN

Backoff[11]是一个经典的语言平滑模型,在n-gram模型中,高阶数据可以更好地利用历史信息,但同时也会面临数据稀疏的问题,此时如果相对低阶的语法出现频率更高,那么结果会更可靠。Backoff模型的原理为:首先,为作为输入数据参与模型训练的n元语法(对应为n阶RNN模型的所有输入序列)构建频率字典;然后,在此基础上为n元语法的出现频率设置一个阈值threshold_n(1元语法无须设置),对于子段$x_1x_2\cdots x_t$,利用$S_{p,q}$代表子段从位置p到q的子串,$p(x_l)$表示在$S_{1,l-1}$后出现x_l的概率($l>2$),在计算$p(x_l)$时Backoff算法会寻找一个最小的p,保证$S_{p,l-1}$的出现频率大于等于阈值threshold_$l-$p;最后,将$S_{p,l-1}$作为RNN_{l-p}的输入序列,得到下一位预测字符及其对应的生成概率。通过利用稀疏的高阶语法中的低阶语法对其进行平滑,这样可以利用给定历史信息选择最可靠的模型来提供更好的预测结果。

基于上述思想,采用算法1来生成长度大于4的子段:首先,遍历长度为i的待生成子串频率字典dict_i(键item为当前待生成子串,值为该子串对应的生成概率);其次,根据训练数据频率字典dict_n确定长度为n的语法的频率阈值(dict_n的键为n元语法,值为该n元语法在训练集中的出现频率);然后,从item[-s:]开始,找到满足阈值的最大子串win,其长度len是RNN模型的阶数,win作为RNN_{len}的输入,得到item的下一位预测字符字典dict_pred(字典的键值对分别为预测字符pred以及对应的生成概率),当生成字符为结束符时,即生成一个完整的字符串,更新dict_generated_i,否则更新dict_i+1,并将其作为长度为i+1的待生成子串;最后,统一对同一长度的已生成子段的概率进行归一化。

特别地,在初始阶段生成长度为5的口令子段时,需要以训练数据频率字典dict_4为初始输入数据得到长度为5的待生成子串字典dict_5后,再将其作为下一步骤的输入数据。

算法1 基于Backoff-RNN的口令字符子串生成算法

输入 训练数据频率字典dict_n($n\in[1,s]$),生成子段长度为i待生成的口令子串频率字典dict_i($i>4$)

输出 已生成的长度为i的子段频率字典dict_generated_i,待生成的长度为i+1的子串频率字典dict_i+1

```
1.FOR item IN dict_i:
2.IF dict_s[item[-s:]]≥threshold_s:
3.win ← item [-s:]
4.ELSE IF dict_(s-1)[item[-(s-1):]]≥threshold_(s-1):
5.win ← item [-(s-1):]
6.反复执行判断,直到得到满足条件的子串
7.updateSubString(win,item)
8.END FOR
```

```
9.def updateSubString(win,item):
10.len = win.length
11.dict_pred_freq ← RNN_len(win)
12.FOR preq IN dict_pred:
13.poss = dict_i[item] × dict_pred[pred]
14.IF pred == <EOS>:
15.dict_generated_i[item] = poss
16.ELSE
17.dict_i+1[item+pred] = poss
18.END FOR
```

3.3.2　改进的概率计算方法

传统PCFG方法是一种典型的基于概率的方法，使用统计学方法从训练集中学习口令的结构分布以及各长度子段的分布，通过概率来刻画每一条口令的分布规律，在构造口令猜测字典时按照概率递减的顺序枚举生成的口令，确保在更少的猜测次数下猜出尽可能多的口令。

最优攻击者生成的猜测集应与测试集完全相同，即按口令概率递减排序后的猜测字典应与按出现频次递减排序后的测试集一致。因此，猜测方法给出的概率值可在最大程度上反映该口令在口令集中的真实频次，这样才能确保基于概率的方法能够有效地根据口令概率加速口令破解。

传统PCFG方法的概率计算规则是将口令结构概率和各子段的概率累乘，然而这种计算方法会导致概率计算不平衡的问题，即结构越长的口令在概率连乘的情况下概率值会越小，例如结构长度为3的口令生成概率普遍会比结构长度为2的口令低一个数量级，从暴力破解方法的实现思路来看，一条口令的长度或结构长度越长，其猜测难度也会相应增大，但根据上述所分析的基于概率的猜测方法的概率生成目的来看，对于一条结构较长的口令A，如果其在口令集中出现的频次高，那么算法也应为其赋予一个较高的概率估计p(A)，同样地，若口令B的结构长度仅为1，但其出现频次更低，那么攻击算法应为其赋予一个比p(A)更低的概率估计。以Yahoo数据集中的5个口令为例，它们在数据集中的出现次数和生成概率如表5所示，可以看出按照此概率计算方法，口令1~3虽然在数据集中出现的次数更多，但由于其口令结构长度均大于口令4和5，其生成概率反而更低。针对上述问题，笔者认为仅使用口令结构概率p(S)足以反映出口令结构分布对整条口令生成概率的影响，在将各子段的概率进行连乘时，需要按照口令结构长度对其进行标准化处理。

表5　部分口令在Yahoo数据集中的出现次数及其生成概率

Table 5　The number of occurrences and their generation probability of some passwords in the Yahoo dataset

序号	口令	出现数/次	口令结构	结构长度	生成概率
1	abc123	250	L_3D_3	2	$p(S=L_3D_3)\times p(L_3=\text{abc})\times p(D_3=123)=3.01\times10^{-5}$
2	abcd1234	71	L_4D_4	2	$p(S=L_4D_4)\times p(L_4=\text{abcd})\times p(D_4=1234)=3.78\times10^{-6}$
3	qwerty123	51	L_6D_3	2	$p(S=L_6D_3)\times p(L_6=\text{qwerty})\times p(D_3=123)=1.88\times10^{-5}$
4	sophie	40	L_6	1	$p(S=L_6)\times p(L_6=\text{sophie})=1.02\times10^{-4}$
5	redsox	32	L_6	1	$p(S=L_6)\times p(L_6=\text{redsox})=9.86\times10^{-5}$

困惑度是衡量句子好坏的指标，对于句子$S_1=w_1w_2\cdots w_n$在uni-gram语言模型下，计算规则如式(3)所示，当困惑度越小时，生成句子的概率越大，语言模型性能越好。

$$P_{\text{Perplexity}}(S_1)=\prod_{i=1}^{n}p(w_i)^{-\frac{1}{n}} \tag{3}$$

在困惑度计算方法的基础上，对基于PCFG的口令概率生成方法进行改进。为了使计算结果直接反映出生成口令的概率大小，在计算时不对概率取倒数，同时为了加速计算以及避免概率连乘导致数值过小而造成浮点数向下溢出的问题，将式(3)通过对数的形式进行转化，计算规则如下：

$$p(S_1)=p(\text{struct})\times 2^{\frac{1}{n}\sum_{i=1}^{n}\text{lb}\,p(w_i)} \tag{4}$$

其中：n表示分割出的子段个数；$p(w_i)$表示第i个子段的概率；$p(S_1)$表示整条口令的生成概率；p(struct)表示口令结构的出现概率。

4　实验与结果分析

4.1　实验设置

实验在Windows 10操作系统下执行，处理器为Intel® Core™ i7-1065G7 CPU，程序编码使用Python 3.6.10及Tensorflow 1.14.0。在实验中网络模型的层数为2，优化器为Adam，在训练时学习率初始化为5×10^{-3}，损失函数为categorical_crossentropy。

4.2　实验场景与评价指标

从传统PCFG方法的实现思路来看，口令结构和子段长度的命中率是影响其攻击效果的重要因素。本节分别设计实验验证所提方法在生成口令结构和子段时相较于其他方法的优势，同时设置以下4种实验场景：

1)将12306数据集和CSDN数据集组成交叉数据集12306_CSDN，随机选择80%作为训练集，20%作为测试集，进行漫步口令攻击，验证所提方法在漫步口令猜测中针对中文语言环境口令的攻击能力。

2)将Yahoo数据集和Rockyou数据集按场景1的规则组合,验证所提方法在漫步口令猜测中针对英文语言环境口令的攻击能力。

3)与PR模型[24]实验数据保持一致,使用CSDN数据集,将其中80%的口令作为训练集,10%的口令作为测试集,10%的口令作为验证集,并在生成5×10^7个口令下将所提方法与PR模型[24]在相同场景下进行对比,验证所提方法的优越性。

4)使用12306数据集,随机抽取80%数据作为训练集,剩下的20%作为测试集,结合个人信息对用户进行定向口令攻击,验证所提方法在定向口令猜测中的攻击性能。

在实验中将命中率P作为评价指标,设T为测试集口令集合,G为猜测方法所生成的候选口令集合,得到:

$$P=\frac{|G\cap T|}{|T|} \tag{5}$$

4.3 结果分析

4.3.1 口令结构生成方法的性能比较

由于12306数据集包含了用户个人信息,可最大程度地反映出口令结构,因此使用此数据集对口令结构的命中率进行实验。使用传统PCFG、RNN、LSTM以及改进生成对抗网络的seqGAN[25]模型来对比不同方法在生成口令结构时的命中率,使用seqGAN代替传统GAN模型进行口令生成是因为seqGAN模型通过引入强化学习,相较于传统GAN模型更适用于文本生成任务。实验结果如图4所示,其中,4阶RNN(RNN_4)及LSTM(LSTM_4)的命中率最高,其他阶数则不做展示。

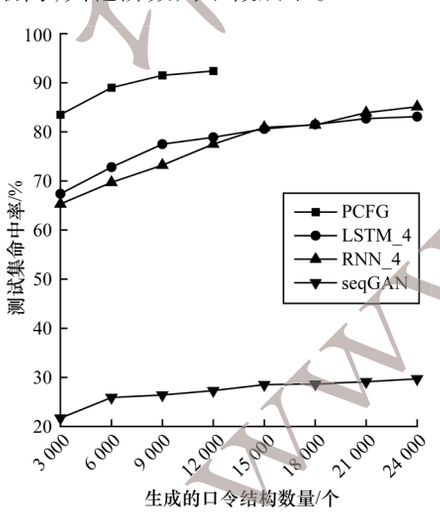

图4 不同口令结构生成方法的性能比较

Fig.4 Performance comparison of different password structure generation methods

在训练集中不重复的口令结构数量为11 857个,由图4可以看出,传统PCFG方法表现优秀,命中率始终高于另外两种方法,对测试集中口令结构的命中率达到92.4%,当生成相同数量的口令结构时,LSTM_4的命中率为78.9%,RNN_4的命中率为77.5%,seqGAN的命中率仅为27.3%。考虑到传统PCFG方法可生成的口令结构数量有限,在实验中继续基于RNN_4、LSTM_4和seqGAN进行口令结构生成,当生成24 000个口令结构时,RNN_4对测试集的命中率为85.1%,LSTM_4稍低,seqGAN的命中率仍仅有29.7%,虽然命中率会随着生成的口令结构个数的增加而不断提高,但口令结构生成方法还要考虑生成效率的问题,应尽可能在更少的猜测数下击中更多的口令。可见,相较于其他3种方法,传统PCFG方法在口令结构的生成中无论是命中率还是效率都更有优势。

从口令结构生成方法的实验结果来看,seqGAN性能较RNN或LSTM差距较大,而RNN和LSTM的命中率极为接近,但前者的训练效率明显高于后者,因此在下文实验中不再将seqGAN和LSTM参与对比。另外,通过口令结构生成方法的实验结果还可以看出,口令集中结构长度多数为1~4,该结果也证明了传统PCFG方法在生成短文本序列时的优势。

4.3.2 模型参数选择与口令子段命中率分析

在12306数据集上设计实验验证Backoff-RNN模型是否能提升对字符长度大于4的子段的猜测命中率。对Backoff-RNN模型中2个重要参数的选择进行纵向对比实验:1)变阶模型的最高阶数s;2)子串的出现频率阈值,以数字子段为例,实验结果如表6所示,其中最优值用加粗标示。由表6可以看出,当最高阶数为5且阈值为80%时,模型性能最优。

表6 模型参数选择

Table 6 Model parameter selection

s	出现频率阈值/%	命中率/%					
		子段长度为5	子段长度为6	子段长度为7	子段长度为8	子段长度为9	子段长度为10
4	70	43.8	42.1	21.8	42.4	**17.3**	13.2
	80	46.1	47.4	22.7	42.5	16.3	14.6
	90	45.4	46.3	20.9	43.7	15.9	14.9
5	70	46.7	49.9	**26.1**	40.9	14.3	13.7
	80	**47.6**	**54.3**	24.3	**45.7**	16.4	**14.8**
	90	43.5	51.3	24.5	45.5	15.4	12.1

基于上述所选参数,将Backoff-RNN与传统PCFG方法、RNN_4、Markov_4进行横向对比,同时为突破传统PCFG方法生成口令子段长度的限制,对实验中其他3种方法生成的口令子段长度均在传统PCFG方法所提供的口令子段长度的基础上增加20%,实验结果如图5所示。由图5可以看出:对于长度为1~4的数字子段,传统PCFG方法保持优势;对于长度大于4的数字子段,传统PCFG方法普遍性能不足,RNN_4及Markov_4虽然相较于前者性能稍有提高,但表现不稳定,Backoff-RNN无论是在命中率还是在稳定性上均优于对比方法。

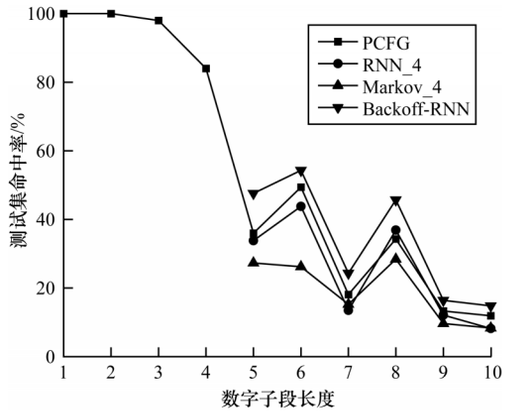

图5　不同口令子段生成方法的性能比较

Fig.5　Performance comparison of different password substring generation methods

图6统计了不同长度数字子段在训练数据中出现的次数,可以看出长度为6~8的子段在训练集中的占比最高,因此使用Backoff-RNN提高长序列口令子段的命中率,对于提高模型整体攻击效果具有重要意义。

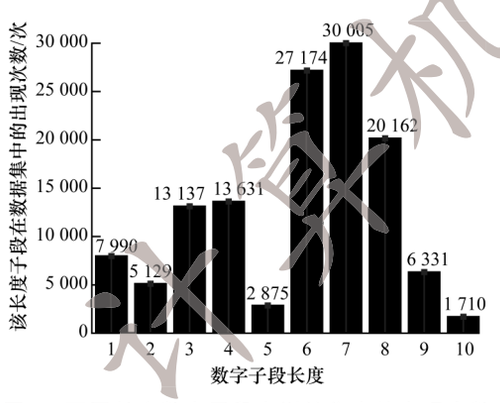

图6　不同长度数字子段在训练集中的出现次数

Fig.6　The number of occurrences of numeric substrings of different lengths in training set

4.3.3　漫步口令攻击命中率分析

通过实验验证所提方法在实验场景1~3下的攻击命中率。在前2个实验场景下,将所提方法与传统PCFG、Markov_4及RNN_4方法的表现性能进行对比。另外,通过设置消融实验进一步验证3个改进方法对PCFG算法命中率提升的有效性。消融实验中的对比方法设置具体如下:

1)Model_1,在传统PCFG方法上增加改进的口令划分策略。

2)Model_2,在Model_1的基础上使用RNN_4来生成长度大于4的字符子段。

3)Model_3,在Model_1的基础上使用Backoff-RNN来生成长度大于4的字符子段。

4)Model_4,所提方法,在Model_3的基础上使用改进的口令概率计算规则。

在前2种实验场景下的漫步口令攻击结果如图7所示。

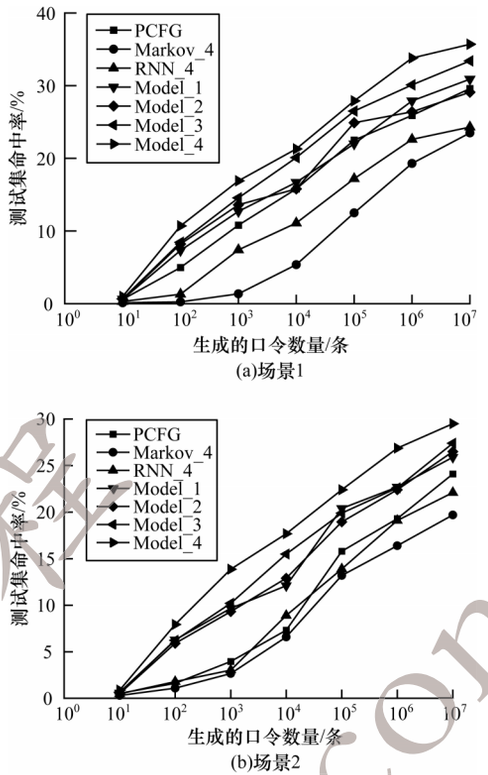

图7　不同实验场景下的漫步口令攻击结果

Fig.7　Trawling password attack results in different experimental scenarios

由图7可以看出:

1)Markov_4相较于其他方法综合表现较差,其性能虽然随着口令生成个数的增加而提升,但仍存在差距。

2)传统PCFG方法相较于Markov_4性能稍有提升,尤其是在生成样本数量较少的情况下。

3)RNN_4的综合性能与传统PCFG方法几乎持平,且仍有提升的空间。

4)Model_1和Model_2相较于传统PCFG方法的性能均有所提升。Model_3的综合表现相较于前两者更稳定且命中率更高,当生成10^7条口令时:在场景1中,Model_3的命中率相较于传统PCFG方法提升了12.8%,相较于Markov_4提升了42.1%;在场景2中,Model_3的命中率相较于传统PCFG方法提升了13.7%,相较于Markov提升了39.1%。该结果证明了Backoff_RNN的引入更加稳定地提升了传统PCFG方法的命中率。

5)Model_4在初始生成阶段的命中率有明显提升:当生成10^3条口令时,对于场景1和场景2,命中率相较于Model_3提升了6.9%及7.6%;当生成10^7条口令时,对于场景1和场景2,Model_4命中率分别为35.7%和29.5%,传统PCFG方法的命中率分别为29.6%及24.1%,Markov_4的命中率分别为23.5%及19.7%。因此,Model_4的命中率相较于传统PCFG方法分别提升了20.6%与22.4%,相较于Markov_4提升了51.9%与49.7%。由于在实验中严格按照口

令概率降序的顺序进行猜测，并且随着猜测数量的不断增加，Model_4命中率始终高于Model_3，因此该结果说明了相较于传统概率计算方法，改进后的概率计算规则能保证概率降序结果，并且对口令分布的刻画能力更优，更能反映出口令在口令集中的真实频次。

为了与当前最新研究进展进行对比，参照文献[24]中已有数据，在相同实验条件下，当生成5×10^7条口令时，所提方法命中率相较于相同场景下的PR模型提升了2.6个百分点。

综上所述，相较于传统方法，所提方法在口令结构分割、子段生成以及口令生成概率计算上均有一定的性能提升，并且提高了漫步口令攻击算法的命中率。

4.3.4 定向口令攻击命中率分析

除漫步口令攻击外，在场景4下验证所提方法在定向口令攻击中的实验性能，并与Personal-PCFG[18]与TarGuess-I[19]模型进行性能对比，如表7所示。对于定向攻击，先提取用户的个人信息字段，再将其作为原始数据对模型进行训练，同时在生成口令时保留个人信息字段，不再对其进行填充。由表7可以看出，在生成100条口令时，所提方法提高了在生成小规模数据时实施定向口令攻击的命中率，相较于Personal-PCFG提升了53.9%，相较于TarGuess-I提升了2.8%。

表7 定向口令攻击结果

Table 7 Targeted password attack results %

猜测方法与模型	命中率
Model_4	21.4
TarGuess-I模型	20.8
Personal-PCFG模型	13.9

5 结束语

本文设计一种基于Backoff-RNN与概率平衡的PCFG口令猜测方法。该方法通过对口令进行更细粒度的结构划分，提取出用户口令更深层次的结构信息。使用Backoff-RNN模型生成长度大于4的字符子段，动态平衡模型拟合问题。对传统PCFG方法中生成口令概率的计算方法进行改进，将困惑度融入口令概率计算规则中，使得口令概率更能体现出口令在口令集中的真实分布规律。实验结果表明，相较于基于PCFG、Markov、RNN等的口令猜测方法，所提方法在漫步攻击和定向攻击中对测试集均具有更高的命中率。后续将提高口令生成算法与概率计算的效率，同时在该猜测方法的基础上构建口令强度评估机制，更好地指导用户创建安全可靠的口令。

参考文献

[1] WANG P, WANG D, HUANG X. Advances in password security[J]. Computer Research and Development, 2016, 53(10): 2173-2188.

[2] BONNEAU J, HERLEY C, VAN OORSCHOT P C, et al. Passwords and the evolution of imperfect authentication[J]. Communications of the ACM, 2015, 58(7): 78-87.

[3] WEIR M, AGGARWAL S, DE MEDEIROS B, et al. Password cracking using probabilistic context-free grammars[C]// Proceedings of the 30th IEEE Symposium on Security and Privacy. Washington D. C., USA: IEEE Press, 2009: 391-405.

[4] HRANICKÝ R, LIŠTIAK F, MIKUŠ D, et al. On practical aspects of PCFG password cracking[M]. Berlin, Germany: Springer, 2019.

[5] 章梦礼，张启慧，刘文芬，等. 一种基于结构划分及字符串重组的口令攻击方法[J]. 计算机学报，2019，42(4)：913-928.
ZHANG M L, ZHANG Q H, LIU W F, et al. A method of password attack based on structure partition and string reorganization[J]. Chinese Journal of Computers, 2019, 42(4): 913-928. (in Chinese)

[6] 罗敏，张阳. 一种基于姓名首字母简写结构的口令破解方法[J]. 计算机工程，2017，43(1)：188-195，200.
LUO M, ZHANG Y. A password cracking method based on name initials shorthand structure [J]. Computer Engineering, 2017, 43(1): 188-195, 200. (in Chinese)

[7] NARAYANAN A, SHMATIKOV V. Fast dictionary attacks on passwords using time-space tradeoff [C]// Proceedings of the 12th ACM Conference on Computer and Communications Security. New York, USA: ACM Press, 2005: 364-372.

[8] MA J, YANG W N, LUO M, et al. A study of probabilistic password models[C]//Proceedings of IEEE Symposium on Security and Privacy. Washington D. C., USA: IEEE Press, 2014: 689-704.

[9] 安亚巍，罗顺，朱智慧. 基于马尔可夫链的口令破解算法[J]. 计算机工程，2018，44(11)：119-122.
AN Y W, LUO S, ZHU Z H. Password cracking algorithm based on Markov chain[J]. Computer Engineering, 2018, 44(11): 119-122. (in Chinese)

[10] MELICHER W, UR B, SEGRETI S M, et al. Fast, lean, and accurate: modeling password guessability using neural networks[C]//Proceedings of the 25th USENIX Conference on Security Symposium. New York, USA: ACM Press, 2016: 175-191.

[11] KATZ S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1987, 35(3): 400-401.

[12] XIA Z Y, YI P, LIU Y Y, et al. GENPass: a multi-source deep learning model for password guessing [J]. IEEE Transactions on Multimedia, 2019, 22(5): 1323-1332.

[13] ZHANG Y, XIAN H Q, YU A M. CSNN: password guessing method based on Chinese syllables and neural network[J]. Peer-to-Peer Networking and Applications, 2020, 13(6): 2237-2250.

[14] NAM S, JEON S, KIM H, et al. Recurrent GANs password cracker for IoT password security enhancement [J].

Sensors，2020，20(11)：3106.
[15] PASQUINI D，GANGWAL A，ATENIESE G，et al. Improving password guessing via representation learning [C]//Proceedings of IEEE Symposium on Security and Privacy. Washington D. C.，USA：IEEE Press，2021：1382-1399.
[16] VERAS R，COLLINS C，THORPE J. On the semantic patterns of passwords and their security impact [C]// Proceedings of 2014 Network and Distributed System Security Symposium. San Diego，USA：Internet Society，2014：1-10.
[17] HOUSHMAND S，AGGARWAL S，FLOOD R. Next Gen PCFG password cracking [J]. IEEE Transactions on Information Forensics and Security，2015，10(8)：1776-1791.
[18] LI Y，WANG H N，SUN K. A study of personal information in human-chosen passwords and its security implications [C]//Proceedings of the 35th Annual IEEE International Conference on Computer Communications. Washington D. C.，USA：IEEE Press，2016：1-9.
[19] WANG D，ZHANG Z J，WANG P，et al. Targeted online password guessing：an underestimated threat [C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York，USA：ACM Press，2016：1242-1254.
[20] HITAJ B，GASTI P，ATENIESE G，et al. PassGAN：a deep learning approach for password guessing [M]. Berlin，Germany：Springer，2019.
[21] 周环，刘奇旭，崔翔，等. 基于神经网络的定向口令猜测研究[J]. 信息安全学报，2018，3(5)：25-37.
ZHOU H，LIU Q X，CUI X，et al. Research on targeted password guessing using neural networks[J]. Journal of Cyber Security，2018，3(5)：25-37. (in Chinese)
[22] XU L Z，GE C，QIU W D，et al. Password guessing based on LSTM recurrent neural networks [C]//Proceedings of IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). Washington D. C.，USA：IEEE Press，2017：785-788.
[23] LIU Y Y，XIA Z Y，YI P，et al. GENPass：a general deep learning model for password guessing with PCFG rules and adversarial generation [C]//Proceedings of IEEE International Conference on Communications. Washington D. C.，USA：IEEE Press，2018：1-6.
[24] 汪定，邹云开，陶义，等. 基于循环神经网络和生成式对抗网络的口令猜测模型研究[J]. 计算机学报，2021，44(8)：1519-1534.
WANG D，ZOU Y K，TAO Y，et al. Password guessing based on recurrent neural networks and generative adversarial networks[J]. Chinese Journal of Computers，2021，44(8)：1519-1534. (in Chinese)
[25] YU L T，ZHANG W N，WANG J，et al. SeqGAN：sequence generative adversarial nets with policy gradient [C]// Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto，USA：AAAI Press，2017：1-10.

编辑　陆燕菲