·安全技术·　　文章编号：1000-3428(2018)11-0119-04　　文献标志码：A　　中图分类号：TP309

基于马尔可夫链的口令破解算法

安亚巍，罗　顺，朱智慧

（上海通用识别技术研究所，上海 201112）

摘　要：口令破解是电子取证的关键技术之一，为克服口令破解中“长度防火墙”问题，在马尔可夫链模型基础上提出一种口令破解算法。通过统计方法对口令空间进行截断，动态给出对口令字符一步状态转移概率矩阵的估计，模拟口令字符设置的潜在规律，并以此得到下一位口令字符的遍历空间。实验结果表明，与普通暴力破解方式相比，该算法的破解效果得到显著提升。

关键词：马尔可夫链；口令破解；口令空间截断；状态转移概率矩阵；概率估计

中文引用格式：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.

英文引用格式：AN Yawei，LUO Shun，ZHU Zhihui. Password cracking algorithm based on Markov chain[J]. Computer Engineering，2018，44(11)：119-122.

Password Cracking Algorithm Based on Markov Chain

AN Yawei，LUO Shun，ZHU Zhihui

（Shanghai General Recognition Technology Institute，Shanghai 201112，China）

【Abstract】Password cracking is one of the key techniques of electronic forensics. In order to overcome the problem of “length firewall” in password cracking，based on Markov chain，a password cracking algorithm is proposed. The password space is truncated by statistical method，the estimation of one-step state transition probability matrix of password character is given dynamically to simulate the potential law of password character setting，and the traversal space of the next password character is obtained. Experimental results show that，compared with the common violent cracking method，the effect of this algorithm is improved significantly.

【Key words】Markov chain；password cracking；password space truncation；state transition probability matrix；probability estimation

DOI：10.19678/j.issn.1000-3428.0048649

0　概述

随着计算机技术和现代密码学的发展，加密算法越来越成熟，加密工具越来越普遍，并且随着人们信息安全意识的普遍提高，口令密码作为一种应用方便、成本低廉的安全手段被越来越多用户所接受[1]，这给电子取证工作带来挑战[2]。越来越多的科研人员关注如何确保加密数据取证结果的准确性和时效性，而口令破解是电子取证领域主要的研究内容之一。

口令破解的主要方法包括弱口令扫描法、字典破解法、暴力破解法等。其中：弱口令扫描法是指遍历低位数或简单字符空间的口令集；字典破解法是指用英文字典、互联网侧漏暴库密码表等预置密码口令进行破解尝试；暴力破解法是指依次试遍所有可能位数长度的口令字符组合。弱口令扫描法和字典破解法都存在需要进行优化以提高口令覆盖率的问题，暴力破解法受时间和存储空间限制，对具有一定长度的目标口令往往无能为力。

为提高口令破解效率，现有的研究已提出多种破解加速方法。一类方法是利用 GPU[3-6]、FGPA[7]等硬件设备或采用分布式计算环境[8]提高破解运算速度；另一类是寻找口令空间分布的规律[9-11]，在小幅降低命中率条件下大幅缩小穷举空间，实现破解效率的提升。

多数口令的设置都包含了很多人为因素和潜在规律。如在国内某知名网站泄露的口令中，频率前五的口令是纯数字口令，频率前 100 的口令全部由小写字符和数字组成；又如在另一份互联网上泄露的密码表中，其 33% 的密码由 8 个字符组成，19% 的密码由 9 个字符组成，16% 的密码由 6 个字符组成，62% 的密码由小写字母和数字组成等[12-13]。本文通

基金项目：国家科技支撑计划项目(2014BAH41B03)。

作者简介：安亚巍(1978—)，男，工程师、硕士，主研方向为信息安全、数据分析处理、知识工程；罗　顺、朱智慧，工程师。

收稿日期：2017-09-14　　**修回日期：**2017-10-27　**E-mail：**ywan20@163.com

过一步状态转移概率矩阵迭代得到的马尔可夫链，本质上即为从已破解口令中挖掘出的人们在进行口令设置过程中的某些行为习惯或语言使用特性[14]。

本文从字典和已破解口令中挖掘规律，提出一种基于马尔可夫链的口令破解算法。通过对字典或已破解口令的统计分析，实现口令字符一步状态转移概率的动态估计，并得到下一位口令字符的遍历空间和遍历顺序。

1 马尔可夫模型

假设一个状态离散的随机变量$\{X_n \mid n=0,1,\cdots\}$有 m 种状态值$\{s_0,s_1,\cdots,s_m \mid m \leqslant n\}$，若变量 X 在某一时刻 X_i处于状态 $s_{j_i}(0 \leqslant j_i \leqslant m)$的概率只与其在前一时刻 X_{i-1}的状态 $s_{j_{i-1}}$有关，即：

$$P(X_i=s_{j_i} \mid X_{i-1}=s_{j_{i-1}},X_{i-2}=s_{j_{i-2}},\cdots,X_0=s_{j_0}) = P(X_i=s_{j_i} \mid X_{i-1}=s_{j_{i-1}})$$

则称变量 X 具有马尔可夫性，变量 X 的随机过程$\{X_n \mid n=0,1,\cdots\}$为马尔可夫链[15]。相应地，在系统变量中具有马尔可夫性的模型被称为马尔可夫模型(Markov Model，MM)。

马尔可夫性的直观含义是在已知现在状态条件下，过去与将来相互独立。即如果用已知的、到现在为止的所有信息来预测将来，则将来只与现在有关，而与过去无关[16-17]。

若令具有马尔可夫性的变量 X 在时刻 k 处于状态 s_i，在时刻 $k+1$ 处于状态 s_j的概率为 p_{ij}，则一步状态转移概率矩阵表示为：

$$\boldsymbol{P}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

其中，$p_{ij} \geqslant 0$ 且$\sum_{j=1}^{m} p_{ij}=1(i=0,1,\cdots,m)$。变量 X 的一步状态转移过程为：

$$(p_{k+1}(s_0),p_{k+1}(s_1),\cdots,p_{k+1}(s_m)) = (p_k(s_0),p_k(s_1),\cdots,p_k(s_m))\boldsymbol{P}$$

其中，$p_k(s_i)$表示变量 X 在时刻 k 处于状态 s_i的概率。若令 $v(k)=(p_k(s_0),p_k(s_1),\cdots,p_k(s_m))$为时刻 k 的状态概率分布，即有：

$$v(k+1)=v(k)\boldsymbol{P}$$

从而可以得到变量 X 在任意时刻 k 的状态概率分布为：

$$v(k)=v(0)\boldsymbol{P}^k$$

其中，$v(0)$为变量 X 的初始状态概率分布，$v(0)$经过与一步状态转移概率矩阵 $\boldsymbol{P}$ 的 k 次迭代得到 $v(0)\boldsymbol{P}^k$，其即为变量 X 转移 k 步后的状态概率分布预测。

2 基于样本估计的马尔可夫链

2.1 口令字符的空间截断

根据 Shannon 信息理论，熵是用来度量随机变量的不确定性。一个离散随机变量 X，其值域记为 S_x，对 S_x中状态值 $s \in S_x$，其概率分布函数为 $p_s(x)$，则变量 X 的熵为：

$$H(X)=-\sum_{s \in S_x} p_s(x)\mathrm{lb}\,p_s(x)$$

对每一个口令字符进行基于样本的统计，并用频率 $f(\alpha)$给出概率 $p(\alpha)$的近似估计。其中，α 表示口令字符，$f(\alpha)$表示在样本字典或已破解口令中字符 α 的出现频率，$p(\alpha)$表示在设置口令时用到字符 α 的概率。由此得到字典或已破解口令中口令的信息熵为：

$$H=-\sum_{\alpha \in S_x} f(\alpha)\mathrm{lb}\,f(\alpha)$$

Algoritmy 网站列出英文字母的出现频率如表 1 所示。

表 1 Algoritmy 网站英文字母出现频率 %

英文字母	出现频率	英文字母	出现频率
a	8.17	n	6.75
b	1.49	o	7.51
c	2.78	p	1.93
d	4.25	q	0.10
e	12.70	r	5.99
f	2.23	s	6.33
g	2.02	t	9.06
h	6.09	u	2.76
i	6.97	v	0.98
j	0.15	w	2.36
k	0.77	x	0.15
l	4.03	y	1.97
m	2.41	z	0.07

计算得出英文语言的信息熵为 4.176，即英文所传达的信息大概只使用了 $2^{4.176} \approx 18$ 个字符。在英文字母中出现频率最高的 18 位字母，占全部小写字母的 69.23%，覆盖了 94.36% 的字母出现频率。

对暴力破解来说，将口令字符在高频的 18 位字符处作截断，意味着若暴力的口令空间为 8 位小写字符，运算量即从 26^8 减少到 18^8，仅相当于原运算量的 5.28%。

在实际破解工作中，口令字符的状态空间为 95 个可打印字符，包括大小写英文字母、数字、标点符号。通过信息熵的计算来进行口令字符的空间截断，其运算缩减量比单纯英文小写字母效果更显著。若仍以 8 位长度口令为例，对某密码表的统计计算得到其信息熵为 5.70，则其有效字符为 52 位，截断后其计算量仅相当于原运算量的 0.81%。

2.2　一步状态转移概率估计

使用马尔可夫链进行口令破解的关键是对口令字符一步状态转移概率的估计。本文通过统计字典或已破解口令中,口令字符在当前各种状态值的情况下向下一刻各种状态值转移的频率,来对一步状态转移概率进行估计。

对统计字典或已破解口令中所有二元字符进行组合,将其前一位字符看成是当前字符状态值,后一位字符看成是下一刻字符转移状态值,即可通过统计所有二元字符组合的字符间跟随关系得到一步状态转移概率矩阵的估计,表示如下:

$$\begin{array}{c} \quad 0\ 1\ \cdots\ 9 \quad a\ b\ \cdots\ z \quad A\ B\cdots Z \quad `\ @\ \cdots ? \\ \begin{matrix} 0 \\ 1 \\ \vdots \\ ? \end{matrix} \begin{pmatrix} f_{00}\ f_{01}\cdots f_{09} & f_{0a}\ f_{0b}\cdots f_{0z} & f_{0A}\ f_{0B}\cdots f_{0Z} & f_{0`}\ f_{0@}\cdots f_{0?} \\ f_{10}\ f_{11}\cdots f_{19} & f_{1a}\ f_{1b}\cdots f_{1z} & f_{1A}\ f_{1B}\cdots f_{1Z} & f_{1`}\ f_{1@}\cdots f_{1?} \\ \vdots & \vdots & \vdots & \vdots \\ f_{?0}\ f_{?1}\cdots f_{?9} & f_{?a}\ f_{?b}\cdots f_{?z} & f_{?A}\ f_{?B}\cdots f_{?Z} & f_{?`}\ f_{?@}\cdots f_{??} \end{pmatrix} \cong \boldsymbol{P} \end{array}$$

其中,$f_{\alpha\beta}$为字典或已破解口令中二元字符组合 $\alpha\beta$ 的频率,即是口令字符从状态值 α 一步转移到 β 的频率,α、β 为可打印字符。

2.3　马尔可夫链的生成

采用第 2.1 节方法对口令字符空间进行截断。设经过截断后口令字符的状态空间从 95 个可打印字符缩减到 $m+1$ 位,记为$\{\alpha_0,\alpha_1,\cdots,\alpha_m\}$。然后得到口令字符一步状态转移概率矩阵的估计 $\boldsymbol{P}=(p_{ij})_{(m+1)\times(m+1)}$。

对矩阵 $\boldsymbol{P}$ 中每一个行向量$(p_{i0},p_{i1},\cdots,p_{im})$,$i=0,1,\cdots,m$ 进行降序排列,记排序结果为$(p_{ij_0},p_{ij_1},\cdots,p_{ij_m})$,$p_{ij_0}\geqslant p_{ij_1}\geqslant\cdots\geqslant p_{ij_m}$,其中,$\{j_0,j_1,\cdots,j_m\}$为$\{0,1,\cdots,m\}$的一个排列。即在根据口令字符一步状态转移概率矩阵得到当前口令字符为 α_i的情况下,下一位口令字符按可能性从高至低依次为 $\alpha_{j_0},\alpha_{j_1},\cdots,\alpha_{j_m}$。

针对字典或已破解口令的马尔可夫链如图 1 所示。

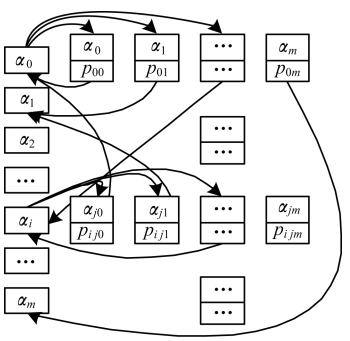

图 1　字典或已破解口令的马尔可夫链

从而得到下一位口令字符的遍历空间$\{\alpha_{j_0},\alpha_{j_1},\cdots,\alpha_{j_m}\}$和遍历顺序$\{j_0,j_1,\cdots,j_m\}$。

2.4　转移矩阵计算

转移矩阵 $\boldsymbol{P}$ 有实时和延时 2 种计算方式。实时计算是指每获得一条口令即进行一次口令空间字符截断计算,同时更新转移矩阵 $\boldsymbol{P}$。实时计算转移矩阵增加了大量的计算开销,而带来的破解性能提升却有限,特别是当口令库规模较大时,入库一条或两条新口令,对口令空间和转移矩阵 $\boldsymbol{P}$ 的影响是非常微弱的。

延时计算是指当入库的新口令达到一定量时(如原口令库的 5%),或者口令字符的频次出现较大变化时(如某一字符新增频次超过 5%),再进行一次口令空间字符截断计算和转移矩阵更新计算。由于延时计算是一次性计算,并且只有当口令库出现较大增量的情况下才会触发,因此,相比于大量的解密运算,延时计算新增加的工作量微乎其微,并极大地保留了算法对破解性能的提升。

3　算法描述

基于马尔可夫链的口令破解算法步骤如下:

步骤 1　统计字典或已破解口令中口令字符的分布。

步骤 2　根据信息熵方法,确定口令空间的截断,得到口令空间的近似。

步骤 3　统计字典或已破解口令中所有二元字符组合,根据二元字符的跟随分布,得到口令字符一步状态转移概率矩阵 $\boldsymbol{P}$ 的估计。

步骤 4　根据口令字符一步状态转移概率矩阵 $\boldsymbol{P}$ 构建马尔可夫链。

步骤 5　根据马尔可夫链,对当前口令字符 α_i 的下一位口令字符,做出遍历空间$\{\alpha_{j_0},\alpha_{j_1},\cdots,\alpha_{j_m}\}$和遍历顺序$\{j_0,j_1,\cdots,j_m\}$的估计。

步骤 6　根据当前口令长度设置,对初始口令字符 α_0,按步骤 5 中方法逐位估计口令空间和顺序,并对口令字符进行遍历。

步骤 7　若步骤 6 中得到正确口令,则退出,否则更改初始口令字符 α_0 为 α_1。

步骤 8　若步骤 7 中得到正确口令,则退出,否则更改当前口令长度设置,并回到步骤 6,若口令长度超限,则退出。

4　测试算例

为验证基于马尔可夫链的口令破解算法的有效性,从网上泄漏的真实用户口令中随机抽取 100 万条作为破解目标设计测试算例。算例以 MD5 算法为例,采用 2 块 AMD Radeon R9 GPU 作为计算设备,对比基于马尔可夫链的破解方式和普通暴力破解方式的破解效率。

以某开源口令破解工具附带密码表为字典,计算口令字符一步状态转移概率矩阵 $\boldsymbol{P}$。该密码表共有 3 399 474 条口令,长度从 1 位到 35 位,其中长度

8 位的口令有 446 739 条。以长度 8 位的口令为样本,统计得到字符分布如表 2 所示。

表 2 某密码表 8 位口令样本字符分布 %

英文字母	出现频率	英文字母	出现频率
e	8.92	r	5.45
a	7.45	t	4.59
i	6.03	u	4.21
s	6.01	l	3.42
o	5.75	k	3.33
n	5.65	⋮	⋮

计算得密码表中口令的信息熵为 4.75,从而将口令空间截断为 $2^{4.75}\approx 29$ 位。也可以根据破解需求进一步放宽对口令空间的截断,如在前 30 位口令字符处进行截断,即{e,a,i,s,o,n,r,t,u,l,k,d,m,c,h,g,p,b,y,v,f,,−,.,′,z,w,j,A,S},其口令字符一步状态转移概率矩阵 $\boldsymbol{P}$ 为:

$$\begin{array}{c} \\ \mathrm{e} \\ \mathrm{a} \\ \vdots \end{array}\begin{array}{c} \begin{array}{ccc} \mathrm{e} & \mathrm{a} & \cdots \end{array} \\ \begin{pmatrix} 2.87\% & 2.54\% & \cdots \\ 2.40\% & 2.47\% & \cdots \\ \vdots & \vdots & \vdots \end{pmatrix} \end{array}$$

设定猜测长度为 7 位~8 位,按前 30 位字符截断,计算得口令猜测空间为 6 779 亿条。采用基于马尔可夫链算法破解,记录实际耗时 118 s,成功破解 246 238 条口令。采用普通暴力破解方式,记录实际耗时 117 s,成功破解 31 条口令。破解测试结果如表 3 所示。

表 3 2 种破解方式的测试结果

方法	破解耗时/s	破解口令数	破解成功率/%
基于马尔可夫链算法	118	246 238	24.623 8
普通暴力破解方法	117	31	0.003 1

从表 3 可以看出,两者比较效果明显,基于马尔可夫链的破解方式能优先遍历更高可能性的字符组合,对口令破解尤其是需要超长破解时间的长口令破解有显著的效果提升。

5 结束语

人们在设置口令过程中的某些行为习惯或语言特性可以用马尔可夫链进行描述。本文提出的基于马尔可夫链的口令破解算法,能够充分利用字典或已破解口令中字符分布、组合分布等特征来提升破解性能。相比暴力破解等传统方法,本文算法具有以下优势:将口令设置的潜在社会工程学特征(人们在设置口令时的某些行为习惯或语言使用特性)量化为口令空间截断和状态转移矩阵,使破解过程指向更有可能性的口令并使计算资源使用更高效;马尔可夫链的使用有效缩小了破解空间,并加速了符合社会工程学规律的口令破解,在同等计算资源条件下,能使口令尽可能快和集中地被破解出来。下一步将继续研究口令中某一字符和与其关联紧密的前两位字符之间关联规律的二阶马尔可夫模型,挖掘实际口令设置规律,以提高口令破解的准确性。

参考文献

[1] FOROUZAN B A. 密码学与网络安全[M]. 马振晗,贾军保,译. 北京:清华大学出版社,2009.

[2] 苏 成. 计算机取证与反取证的较量[J]. 计算机安全,2006(1):67-73.

[3] 翁 捷,吴 强,杨灿群. 基于 OpenCL 加速的 MD5 破解算法[J]. 计算机工程,2011,37(4):119-121.

[4] CHEN R, ZHANG Y, ZHANG J, et al. Design and optimizations of the MD5 crypt cracking algorithm based on CUDA [M]. Berlin, Germany: Springer, 2014: 155-164.

[5] 谢鑫君,罗 顺,杨士华. 基于口令自生成的 GPU 暴力破解优化技术[J]. 信息安全与通信保密,2013(3):82-84.

[6] 乐德广,常晋义,刘祥南. 基于 GPU 的 MD5 高速解密算法的实现[J]. 计算机工程,2010,36(11):154-158.

[7] 李龙谱,斯雪明,张志鸿,等. 在多 FPGA 上实现基于字典的 ZIP 文档口令恢复[J]. 计算机应用与软件,2015,32(6):292-295.

[8] 石志才. 异构平台上协同计算的相关研究[D]. 长沙:国防科学技术大学,2011.

[9] 陈锐浩,邱卫东. 基于神经网络的口令属性分析方法[J]. 微型电脑应用,2015,31(4):45-47.

[10] 张世良. 字符空间优化搜索策略的研究与实现[D]. 广州:华南理工大学,2012.

[11] 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.

[12] JUNIUS. 用户密码薄如纸[EB/OL]. [2017-08-14]. https://jandan.net/2013/05/29/crackers-make-minced.html/page-1.

[13] ZSHKING. 中国网民密码习惯[EB/OL]. [2017-08-14]. http://weibo.com/zshking.

[14] 梁勇勇. 从密码设置的习惯探讨设置安全密码[J]. 丽水学院学报,2006,28(2):58-60.

[15] 马振华. 现代应用数学手册——概率统计与随机过程卷[M]. 北京:清华大学出版社,2001.

[16] 张 超,贾凤亭. Markov 链的组合预测及其应用[J]. 辽宁工程技术大学学报(自然科学版),2011,30(6):963-966.

[17] 何江宏,陈启明. 基于 Markov 链的最优化预测模型及其应用研究[J]. 合肥学院学报(自然科学版),2006,16(1):11-13.

编辑 司淼淼