文章编号：1004-4353(2021)02-0141-05

基于自动机理论的密码匹配方法

姜克鑫， 赵亚慧*， 崔荣一
（延边大学 工学院，吉林 延吉 133002）

摘要：针对用户安全登录问题，提出了一种基于自动机的密码匹配模型. 首先，对于用户任意输入的密码进行同态映射加密；其次，构造出可接受加密密码的自动机——状态数目可变自动机(VNS-DFA)，该自动机不仅能够匹配加密密码，同时还可以输出加密密码的同态原像以及匹配成功的次数；最后，在状态数目可变的自动机上对用户输入的密码进行实验验证表明，用户建立的密码经过同态映射后可全部被该自动机接受，且该自动机的时间复杂度优于传统的 DFA 以及改进的 DFA.

关键词：自动机理论；密码匹配；密码加密；同态映射；同态原像
中图分类号：TP391.41 **文献标识码**：A

Password matching method based on automata theory

JIANG Kexin， ZHAO Yahui*， CUI Rongyi
（*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: For users to log in safely, a password matching model based on automata was proposed. Firstly, we perform homomorphic mapping encryption for any entered password; Then, we construct an automaton that accepts encrypted passwords, namely variable number of states automata (VNS-DFA), the automat a can not only match the encrypted password, but also can output the homomorphic image of the encrypted password and the number of successful matches. Finally, the experiments on the automata with variable number of states show that the user's password can be accepted by the automata after homomorphic mapping, and the time complexity of the automata is better than that of the traditional DFA and the DFA which was improved.

Keywords: automata theory; password match; password encryption; homomorphic mapping; homomorphic preimage

0 引言

密码匹配问题实质上是字符串的匹配问题[1]. 传统的匹配算法主要有基于字符暴力匹配的 BF 算法[2-3]和通过消除主串指针回溯的 KMP 算法[4]，其中 KMP 算法的时间复杂度虽然低于 BF 算法，但 KMP 算法中因存在着相同字符的重复比较，因此其匹配效率仍相对较低. 2019 年，李先祥等[5]提出了 BM 算法. 由于 BM 算法的速度比 KMP 算法快 3～5 倍，因此其受到学者的广泛关注.

1956 年，Moore 等首次提出了有限状态自动机的概念[6]. 随后，学者们利用有限状态自动机对字符串匹配的问题进行了研究. 例如：文献[7]给出了一种高效的有限状态自动机的存储表示方法，并基于这种存储表示方法建立了一种效率高于 KMP 算法的模式匹配算法；文献[8]提出了一种基于自动机的多模式匹配算法(AC 算法)，该算法在匹配失败时能够高效跳转，因此其匹配效率较好. 但目前相关研究

收稿日期：2020－11－05 **基金项目**：延边大学外国语言文学一流学科建设项目(18YLPY13)
* **通信作者**：赵亚慧(1974—)，女，教授，研究方向为模式识别、智能计算、自然语言处理.

中所提出的自动机状态数目都是固定不变的,即仅能匹配特定的输入字符串,因此具有很大的局限性.为此,本文提出了一种新的有限自动机——状态数目可变自动机(variable number of states automata, VNS-DFA),并通过算法分析验证了该自动机的有效性.

1 有限状态自动机

在自动机理论中,自动机都是指抽象的自动机,即是一种能变化和处理信息的离散动态数学模型.自动机可通过形式化语言、状态转移图以及状态转移表3种方式进行描述[9].目前,自动机主要包括确定型有限状态自动机、非确定型有限状态自动机、有穷概率自动机[10]、模糊有穷自动机[11]、下推自动机[12]、图灵机[13]等.其中确定型有限状态自动机是一种控制状态和符号集都有限的自动机,它能够对每个输入的字符做出识别,并保证所输入的字符串都能够到达最终的状态和路径[14].由于确定型有限状态自动机实现简单,且能应用于字符串匹配中,因此本文选用确定型有限状态自动机.

1)有限状态自动机.有限状态自动机 M 可表示为一个五元组的形式[15]:

$$M=(Q,\Sigma,\delta,q_0,F). \tag{1}$$

其中:Q 为状态的非空有穷集合;q 称为 M 的一个状态;Σ 为输入字母表,输入字符串都是 Σ 上的字符串;δ 为状态转移函数,$\delta: Q\times\Sigma\to Q$,对 $\forall(q,a)\in Q\times E$,$\delta(q,a)=p$,表示 M 在状态 q 时读入字母表的字符 a,将状态 q 变成 p,并处理下一个字符;q_0 为 M 的开始状态,$q_0\in Q$;F 为 M 的终止状态,$F\subseteq Q$.

假设公式(1)中的 $Q=\{q_0,q_1\}$,$\Sigma=\{0,1\}$,q_0 为开始状态,$F=\{q_1\}$ 为终止状态,且 $\delta(q_0,0)=q_0$,$\delta(q_0,1)=q_1$,$\delta(q_1,0)=q_0$,$\delta(q_1,1)=q_1$,此时自动机 M 的状态转移图如图1所示,其状态转移表如表1所示.在图1中,一个圆圈代表一个状态,带标签的箭弧表示转移函数,无标签的箭弧指向的圆圈表示初始状态,双圆圈表示终态.在表1中,表的各行对应 M 的状态,表的各列对应输入事件,其中有箭头标注的状态是初始状态,有 * 标注的状态是终态.

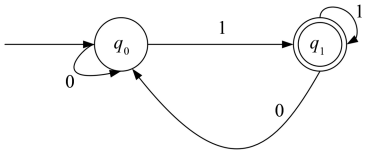

图1 M 的状态转移图

表1 M 的状态转移表

状态	输入事件	
	0	1
→ q_0	q_0	q_1
* q_1	q_0	q_1

由式(1)可知,上述定义的有限状态自动机 M 只能判别输入的字符串是否被自动机接受,但无法给出必要的中间结果和确定自动机 M 所接受的字符串是什么内容.对此,本文在公式(1)中增加了一个输出字母表 Δ 以及输出函数 $g: Q\to\Delta$,且当 $q\in Q$ 时,$g(q)=a$ 表示 M 在 q 状态下输出 a.

2)有限状态自动机接受的语言.自动机接受的所有字符串的集合称为自动机所接受的语言.设 $M=(Q,\Sigma,\delta,q_0,F)$ 是一个自动机,对于 $\forall x\in\Sigma^*$,如果 $\delta(q_0,x)\in F$,则称 x 被 M 接受;如果 $\delta(q_0,x)\notin F$,则称 x 不被 M 接受.有限状态自动机所接受的语言可表示为:

$$L(M)=\{x\mid x\in\Sigma^*\text{且}\delta(q_0,x)\in F\}. \tag{2}$$

2 状态数目可变自动机(VNS-DFA)的构造

为提高用户密码的安全性,本文设计了如下用户注册和用户登录的流程:用户注册时首先设置密码,然后系统对密码进行同态映射加密;用户登录系统时,为防止密码被偷窥,将密码隐藏在所输入的字符串中.为了接受这些隐藏在字符串中的密码,本文构造了一种 VNS-DFA 自动机,只要用户输入的密码被该自动机正确识别即可成功登录.密码设置和 VNS-DFA 自动机识别的过程如图2所示.

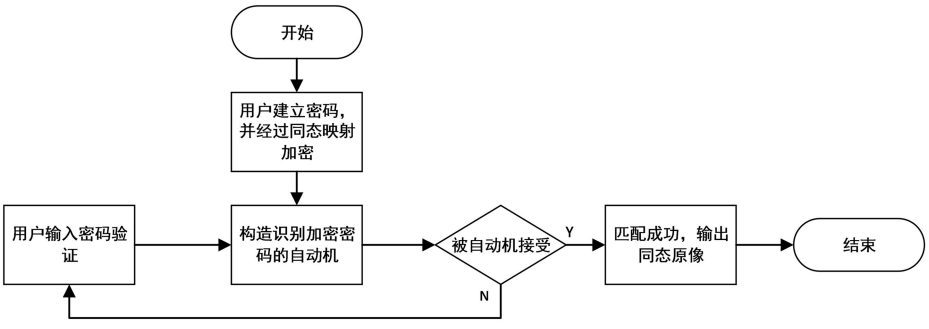

图 2　密码设置和识别的过程

2.1　密码加密的过程

密码加密的过程可以看作是一个同态映射[16].设 Σ 是初始密码的字母表,Δ 是加密之后密码的字母表,$f:\Sigma\to\Delta^*$ 为映射.如果对于 $\forall x,y\in\Sigma^*$ 有 $f(xy)=f(x)f(y)$,则称 f 为从 Σ 到 Δ^* 的同态映射.对于 $\forall L\subseteq\Sigma^*$,$L$ 的同态像为:$f(L)=\bigcup\limits_{x\in L}\{f(x)\}$.对于 $\forall w\subseteq\Delta^*$,$w$ 的同态原像是一个集合,可表示为:

$$f^{-1}(w)=\{x\mid f(x)=w\ \text{且}\ x\in\Sigma^*\}. \tag{3}$$

由式(3)可知,若 w 是经过加密之后的密码,则 w 的原始密码应包含在它的同态原像中.

2.2　自动机的构造及识别

本文设计的状态数目可变自动机(VNS-DFA)可用如下的七元组表示:

$$M=(Q,\Sigma,\Delta,\delta,q_0,F,g). \tag{4}$$

其中:Q,δ,q_0,F 的含义与 DFA 中的含义相同;g 为输出函数,$\exists q\in Q,a\in\Sigma,g(q)=a$,即在满足条件的状态 q 下输出原始字符 a;Σ 为输出字母表;Δ 为 Σ 经过同态映射之后的输入字母表.

根据自动机的结构及原理,本文构造的 VNS-DFA 算法(算法 1)的具体步骤如下所示:

输入:原始串 s,验证串 t.

输出:加密串在验证串中出现的次数.

Step 1　将串 s 进行加密,得到串 $s1$.

Step 2　初始化自动机状态转移矩阵 $\boldsymbol{A}$,并对其元素全部赋值为 0.

Step 3　修改矩阵 $\boldsymbol{A}$,其中 m 为 $s1$ 的长度,0 为初态,n 为终态.具体修改规则如下:

$A[0][s1[0]]=1$;

$A[n][s1[0]]=n+1$;

$A[n+1][s1[1]]=2$;

if $s1[0]==s1[m-1]$;

　$A[n][s1[1]]=2$;

$A[i][s1[0]]=n+1$, $A[i][s1[i]]=i+1$, $i=1,\cdots,m-1$.

Step 4　初始状态 $k=0$,匹配成功次数 $count=0$,待匹配字符位置 $i=0$.

Step 5　如果 $i==n$ 回到 Step 6;否则,跳转至新的状态 $j=A[k][s1[i]]$,并将 j 赋值给 k;

if $j==$特定状态,输出相应数字,回到 Step 5;

if $j==n-1$ 匹配成功,$count+1$,回到 Step 5;

$i=i+1$.

Step 6　结束.

算法 1 中:输入是原始密码 s 和用户验证登录密码 t,输出是 t 在 s 中是否出现及其出现的次数,矩阵 $\boldsymbol{A}$ 为自动机的状态转移矩阵,m 为原始密码经过同态映射后的密码长度,0 为初始状态,n 为终止状态.

Step 3 为算法 1 的核心. 在由 Step 3 建立好的自动机中输入某个字符时,若自动机成功匹配,则跳转至下一状态,否则跳转至初始状态. Step 5 为算法 1 的匹配过程,当且仅当状态跳转为终态时,输入的字符串才能被自动机所接受,同时输出其同态原像.

3 算法分析

3.1 算法的有效性证明

对于任意一个输入串 $a_1a_2\cdots a_n$,根据文献[16]中的定理1(对于任意的 $q\in Q, w\in\Sigma^*, a\in\Sigma$,都有 $\delta(q,wa)=\delta(\delta(q,w),a)$)可以得到如下式子:

$$\delta(0,a_1a_2\cdots a_n)=\delta(\delta(0,a_1a_2\cdots a_{n-1}),a_n)=\delta(\delta\cdots(\delta(0,a_1)\cdots),a_n). \tag{5}$$

根据算法 1 的转移规则($\delta(0,a_1)=1,\cdots,\delta(n-1,a_n)=n$),可将式(5)转变为:

$$\delta(\delta\cdots(\delta(0,a_1)\cdots),a_n)=\delta(\delta\cdots(\delta(1,a_2)\cdots),a_n)=\delta(n-1,a_n)=n. \tag{6}$$

由式(6)可得

$$\delta(0,a_1a_2\cdots a_n)=\delta(\delta\cdots(\delta(1,a_2)\cdots),a_n)=\delta(n-1,a_n)=n. \tag{7}$$

由式(7)可知,自动机在初态时可接受字符串 $a_1a_2\cdots a_n$,并且最后可到达终态. 根据本文提出的算法所构造的识别串 $a_1a_2\cdots a_n$ 的有限状态自动机如图 3 所示.

本文以输入密码 s 为例验证本文构造的自动机能够接受该密码. 假设 $s='12'$,且 s 对应的同态映射为 $f(1)='\text{one}'$, $f(2)='\text{two}'$. 在该假设下匹配该密码的自动机如图 4 所示. 由图 4 可知:当输入的字符串中只要包含 'onetwo',该自动机就可接受该字符串;当自动机匹配到 'one' 时,自动机输出 'one' 对应的同态原像 '1';当自动机匹配到 'two' 时,自动机输出 'onetwo' 对应的同态原像 '12'. 由以上可知,本文提出的自动机能够匹配包含加密密码的字符串,并能够输出用户输入的原始密码,因此本文构造的自动机是有效的.

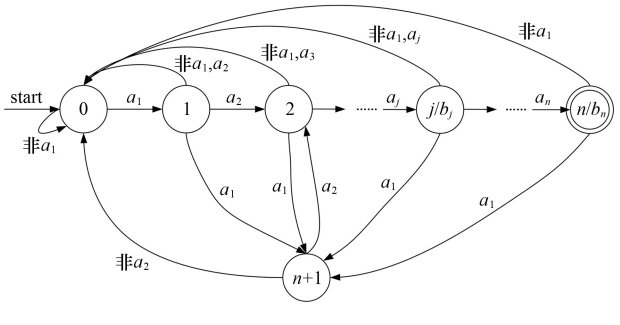

图 3 识别 $a_1a_2\cdots a_n$ 的有限状态自动机

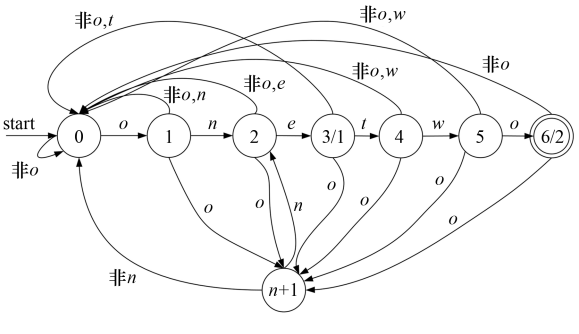

图 4 识别原始串'12'的有限状态自动机

3.2 算法的复杂度分析

自动机算法的复杂度通常包括建立算法的时间复杂度和匹配字符的时间复杂度. 为了验证 VNS-DFA 算法的有效性,本文将 VNS-DFA 算法的时间复杂度、空间复杂度与传统的 DFA 和改进后的 DFA 的时间复杂度、空间复杂度进行了对比,结果如表 1 所示. 实验中,将待匹配串的长度设置为 m,输入串的长度设置为 n,字母长设置为 $|\Delta|$.

表 2 3种不同算法的复杂度

算法	时间复杂度	空间复杂度
DFA	$O(m^3\|\Delta\|)+\Theta(n)$	$O(m\|\Delta\|)$
改进后的 DFA[17]	$O(m\|\Delta\|)+\Theta(n)$	$O(m\|\Delta\|)$
VNS-DFA	$O(m)+\Theta(n)$	$O((m+2)\|\Delta\|)$

由表 2 可知：在时间复杂度上，在输入规模相同的情况下 3 种算法在字符匹配的过程中的时间复杂度均为 $\Theta(n)$，但 VNS-DFA 算法的建立时间最短（复杂度为 $O(m)$）；在空间复杂度上，在输入规模相同的情况下其差别不大. 这是因为 3 种算法都可用二维数组表示状态转移矩阵，因此其大小取决于自动机的状态个数以及输入字母表的长度.

4　结论

研究表明，本文构造的 VNS-DFA 能够匹配任意输入的字符串，且其匹配速率优于传统的 DFA 及改进的 DFA，因此本文方法可以用于密码的识别. 在本文的算法中，当待匹配的字符串长度较大时，会因自动机的状态数目增多而导致耗费更多的内存空间. 因此在今后的研究中，我们将对自动机的状态数目进行改进，以节约内存空间.

参考文献：

[1] 敬茂华. 一种新的自动机构造理论(PFA)[D]. 沈阳：东北大学，2016.

[2] 黄鸿华. 基于 Visual C++的装箱问题的 BF 算法[J]. 电脑知识与技术，2018，14(36)：258-259.

[3] 蔡恒，张帅. 基于 BF 算法改进的字符串模式匹配算法[J]. 电脑编程技巧与维护，2014(22)：14-15.

[4] RAHIM R，ZULKARNAIN I，JAYA H. A Review：search visualization with Knuth Morris Pratt algorithm[C]// IOP Conference Series：Materials Science and Engineering. Medan：IOP，2017，237(1)：012026.

[5] 李先祥，陈思琪，肖红军，等. 基于 SGBM 算法与 BM 算法的三维重建分析[J]. 自动化与信息工程，2019，40(5)：6-12.

[6] 王茁. 基于有限状态自动机的公交车到站时间预测模型[D]. 哈尔滨：哈尔滨工业大学，2012.

[7] 程晓锦，徐秀花. 有限状态自动机及在字符串搜索中的应用[J]. 北京印刷学院学报，2014，22(4)：45-48.

[8] 熊仁都，杨嘉佳，朱广宇，等. PARA-AC：一种基于 AC 自动机的高性能匹配算法[J]. 电子技术应用，2020，46(11)：87-90.

[9] 赵庚兵. 基于自动机理论的软件项目进度监控方法研究[D]. 广州：广东工业大学，2016.

[10] KNAST R. Finite-state probabilistic languages[J]. Information & Control，1972，21(2)：148-170.

[11] BLANCO A，DELGADO M，PEGALAJAR M C. Fuzzy automaton induction using neural network[J]. International Journal of Approximate Reasoning，2001，27(1)：1-26.

[12] DAS S，GILES C L，SUN G Z. Using prior knowledge in an NNPDA to learn context-free languages[J]. Advances in Neural Information Processing Systems，1993，5：65-72.

[13] 宋文，牟行军. 计算的模型：图灵机与 Petri 网[J]. 西华大学学报(自然科学版)，2012，31(3)：1-6.

[14] 罗智勇，杨旭，孙广路，等. 基于马尔可夫的有限自动机入侵容忍系统模型[J]. 通信学报，2019，40(10)：79-89.

[15] 敬茂华，杨义先，汪韬，等. 新颖的正则 NFA 引擎构造方法[J]. 通信学报，2014，35(10)：98-106.

[16] 蒋宗礼，姜守旭. 形式语言与自动机理论[M]. 北京：清华大学出版社，2002：89-91.

[17] THOMAS H C，CHARLES E L，RONALD L R，et al. 算法导论[M]. 殷建平等，译. 3 版. 北京：机械工业出版社，2013：583-588.