文章编号：1004-4353(2022)02-0169-05

基于自动机的用户名合法性检测方法

刘帆[1]， 赵亚慧[1,2]， 崔荣一[2]
（1.延边大学 融合学院，吉林 延吉 133002；2.延边大学 工学院，吉林 延吉 133002）

摘要：针对现有方法检测用户名合法性效率较低的问题，提出了一种基于自动机的用户名合法性检测模型.该模型利用映射函数对用户名字符串进行映射，以此实现由字符串向映射串的转化；利用构造的计数自动机实现对映射串的合法性检测.研究表明，该模型具有检测效率高、性能稳定等优点，因此该方法可应用于用户名合法性的检测中.

关键词：自动机理论；自动机；用户名；合法性检验；同态映射
中图分类号：TP391.41 **文献标识码**：A

Username validity detection method based on automata

LIU Fan[1]， ZHAO Yahui[1,2]， CUI Rongyi[2]
（1. *College of Integration Science*, *Yanbian University*, *Yanji* 133002, *China*；
2. *College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: A user name validity detection model based on automata was proposed to solve the problem of low efficiency and accuracy in user name validity detection. In this model, the user name string is mapped by mapping function to transform from string to mapping string, and the validity of mapping string is detected by counting automata. The research shows that this model has the advantages of simplicity, stability and high detection accuracy, therefore, this method can be applied to websites to detect the validity of registered user names.

Keywords: automata theory; automata; username; legitimacy test; homomorphic mapping

0 引言

用户名合法性检测问题的本质是字符串匹配问题.与传统的字符串匹配不同，合法性检测问题只需保证字符类别匹配成功即可.合法的用户名字符串需同时含有数字、特殊字符、大写字母且其长度需大于 8 位.传统的字符匹配算法主要有暴力匹配(BF)算法[1]、Rabin-Karp(RK)算法[2]、字符串查找(KMP)算法[3]以及正则表达式[4].其中：BF 算法虽然简单，但耗时较大；RK 算法是基于哈希函数的一种对 BF 算法改进的算法，在正常情况下该算法的效率虽然优于 BF 算法，但当哈希产生冲突时该算法的效率仍然较低；KMP 算法是一种采用消除主串指针回溯的方法对 BF 算法进行优化的算法，该算法虽然能够弥补 BF 算法在主针回溯后需重新开始进行比较的缺点，但在重复率很高的文本中其效率低于 BF 算法；正则表达式与上述方法相比，虽然时间效率有所提高，但存在空间复杂度高和易出错等缺点[5].近年来，一些研究者基于上述方法又提出了 BM 算法[6]和 Sunday 算法[7]，这两种方法虽大幅提高

收稿日期：2022-02-18
基金项目：国家语委“十三五”科研项目(YB135-76)；延边大学外国语言文学一流学科建设项目(18ylpy13)
第一作者：刘帆 (2000—)，女，硕士研究生，研究方向为自然语言处理.
通信作者：赵亚慧(1974—)，女，硕士，教授，研究方向为模式识别、智能计算、自然语言处理.

了字符串匹配的效率,但二者过于依赖于辅助数据结构,且预处理时间过长.

1956年,Moore等首次提出了有限状态自动机的概念[8],并在后续的研究中针对字符串匹配问题提出了有限状态自动机算法—Aho-Corasick(AC)算法[9].该算法与上述传统算法相比,只需扫描一遍字符串即可,且其时间复杂度与模式串的规模无关,因此该算法受到学者们的关注.目前利用该方法虽然可以检测用户名字符串中包含的类别,但无法检测用户名长度,因此其应用性受到一定限制.为此,本文提出了一种新有限状态自动机,并通过分析验证了该自动机的有效性.

1 有限状态自动机

传统的有限状态自动机由一个五元组组成[10]:

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

其中:Q 是包含自动机中所有状态的非空集合;Σ 是字母表,自动机接收的任意字符均为该集合元素;δ 是状态转移函数,其中 $\delta:Q\times\Sigma\to Q$,对 $\forall(q,a)\in Q\times\Sigma$,$\exists p\in Q$ 有 $\delta(q,a)=p$,它表示自动机 M 在状态 q 时读入字符 a 之后,状态 q 随即转向状态 p;q_0 为 M 的初始状态,$q_0\in Q$;F 为终止状态集合,$F\subseteq Q$,$\forall q\in F$,q 称为 M 的一个终态.

对于 $\forall x\in\Sigma^*$,若有 $\delta(q_0,x)\in F$,则称 x 被 M 所接受;对于 $\forall x\in\Sigma^*$,若有 $\delta(q_0,x)\notin F$,则称 x 不被 M 所接受[11].所有可以被 M 接受的 x 的集合称为 M 可接受的语言,表示为:

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

假设 $Q=\{q_0,q_1,q_2\}$,$\Sigma=\{0,1\}$,$F=\{q_2\}$,$\delta(q_0,0)=q_1$,$\delta(q_0,1)=q_0$,$\delta(q_1,0)=q_2$,$\delta(q_1,1)=q_0$,$\delta(q_2,0)=q_2$,$\delta(q_2,1)=q_0$,则此时的自动机 M 如图1所示,其状态转移函数如表1所示.在图1中,单圆圈代表自动机中的中间状态,双圆圈代表终止状态,标有字符的有向箭头代表转移函数,箭头S所指向的单圆圈代表初始状态.

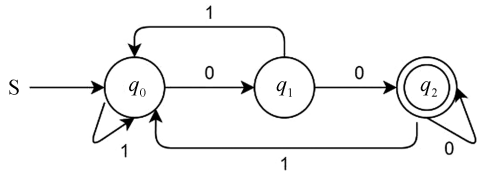

图1 M 的状态转移图

表1 M 的状态转移函数

	状态	输入字符	
		0	1
初始	q_0	q_1	q_0
	q_1	q_2	q_0
终止	q_2	q_2	q_0

2 自动机模型的构建

为了便于检测用户名的合法性,本文设计了一种可计数的有限状态自动机,其检测流程如下:①用户提交需检测的用户名.②系统利用同态映射对用户名进行字符分类(属于同一类的字符被映射为同一个字符,同时用户名字符串映射为对应类别字符串);若某类别字符串可以被该自动机正确识别,则对应的用户名字符串为合法,否则为不合法.由于自动机每次只能输出一个字符,不能直接输出用户名是否合法,因此本文设计了一个解码函数.该解码函数可根据自动机最后的状态输出用户名不合法的原因.用户名映射、自动机识别以及状态解码的过程如图2所示.

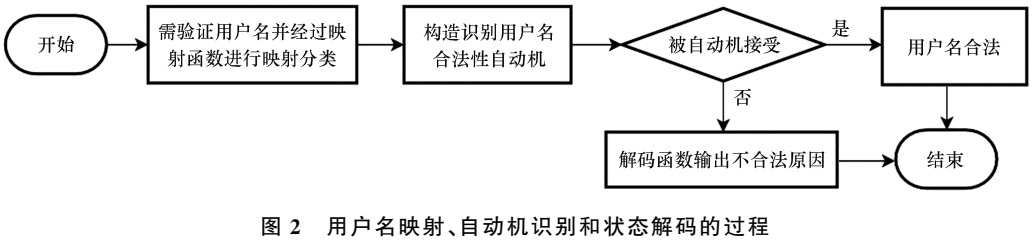

图2 用户名映射、自动机识别和状态解码的过程

2.1　映射函数和解码函数

用户名字符分类是一个同态映射过程. 设 D 是初始用户名字符串，E 是分类字符串，$f:D\to E$ 为映射，对于 $\forall d_1,d_2\in D$ 有 $f(d_1d_2)=f(d_1)f(d_2)$，则称 f 是 D 到 E 的同态映射. 本文将用户名字符分类的映射函数定义为：

$$f(x)=\begin{cases}1, & x \text{ 是数字};\\ a, & x \text{ 是小写字母};\\ T, & x \text{ 是特殊字符};\\ A, & x \text{ 是大写字母}.\end{cases}\tag{3}$$

其中 x 是原始用户名字符串中的字符.

输出不合法原因的解码函数是一个映射过程，即当 $\forall a\in A$，$\exists b\in B$，且有 $h(a)=b$，称 h 是 A 到 B 的映射. 本文将解码函数定义为：

$$h(x)=\begin{cases}\text{“用户名缺少数字、特殊字符以及大写字母，不合法”}, & x==\text{'}q_0\text{'};\\ \text{“用户名缺少特殊字符和大写字母，不合法”}, & x==\text{'}q_1\text{'};\\ \text{“用户名缺少数字和大写字母，不合法”}, & x==\text{'}q_2\text{'};\\ \text{“用户名缺少数字和特殊字符，不合法”}, & x==\text{'}q_3\text{'};\\ \text{“用户名缺少数字，不合法”}, & x==\text{'}q_4\text{'};\\ \text{“用户名缺少特殊字符，不合法”}, & x==\text{'}q_5\text{'};\\ \text{“用户名缺少大写字母，不合法”}, & x==\text{'}q_6\text{'};\\ \text{“用户名长度不够，不合法”}, & x==\text{'}q_8\text{'}.\end{cases}\tag{4}$$

其中，x 是指自动机在接收一个字符串后跳转到的最终状态.

2.2　自动机模型的构建

本文建立的自动机模型 G 由一个六元组组成：

$$G=(Q,\Sigma,\delta,q_0,F,n).\tag{5}$$

其中：Q、Σ、δ、q_0、F 与有限状态自动机的定义一致；n 为计数器（自动机每接受 1 个字符时，n 减 1）. G 的状态转移函数如表 2 所示. 本文算法的具体步骤如下：

表 2　G 的状态转移函数

状态	输入字符			
	a	1	T	A
q_0	q_0	q_1	q_2	q_3
q_1	q_1	q_1	q_6	q_5
q_2	q_2	q_6	q_2	q_4
q_3	q_3	q_5	q_4	q_3
q_4	q_4	q_7	q_4	q_4
q_5	q_5	q_5	q_7	q_5
q_6	q_6	q_6	q_6	q_7
q_7	q_7	q_7	q_7	q_7

输入：字符串 s

输出：用户名是否合法

Step1　对字符串 s 进行映射，得到映射串 s'；

Step2　根据状态转移函数表初始化状态转移字典 trans，同时将 n 初始化为 8；

Step3　设初始状态为 q_0，即设待匹配字符串的首位为 $i=0$；

Step4　读完字符串时，若 $n\leqslant 0$ 时则算法结束，否则跳转到 Step5；

Step5　依据最终状态进行解码并输出字符中不合法的原因，算法结束.

该算法中，输入是需验证合法性的原始用户名 s；输出是该用户名 s 是否合法. 若合法，不输出错误；若不合法，输出不合法并给出原因. 映射串 s' 是输入字符串 s 经函数 $f(x)$ 同态映射得到的分类字符串，字典 trans 是依据表 2 进行初始化的状态转移函数，n 赋值为 8 表示合法用户名至少是 9 位字符；end 表示自动机为终态，$end\in F$.

3 算法分析

3.1 算法有效性证明

对于任意的一个输入字符串 $a_1a_2\cdots a_n$，根据文献[10]中的定理 1 可知：对于 $\forall q\in Q$，$\omega\in\Sigma^*$，$a\in\Sigma$，有 $\delta(q,\omega a)=\delta(\delta(q,\omega),a)$. 由此可得：

$$\delta(q_0,a_1a_2\cdots a_n)=\delta(\delta(q_0,a_1a_2\cdots a_{n-1}),a_n)=\delta(\delta(\cdots\delta(q_0,a_1)\cdots),a_n). \tag{6}$$

再根据表 2 可得：

$$\delta(q_0,a_1a_2\cdots a_n)=\delta(\delta(q_0,a_1a_2\cdots a_{n-1}),a_n)=\delta(\delta(\cdots\delta(q_0,a_1)\cdots),a_n)=end. \tag{7}$$

依据计数器的定义，此时接收到的字符串 $a_1a_2\cdots a_n$ 的值为：

$$n=8-|a_1a_2\cdots a_n|. \tag{8}$$

公式(8)中的 $|a_1a_2\cdots a_n|$ 表示字符串 $a_1a_2\cdots a_n$ 的长度. 当式(8)中求得的 n 大于 0 时，自动机在 end(终态)处跳转至下一状态，即表示用户名不合法. 因此，只有当自动机最后到达的状态是终态时，该用户名才是合法的，否则为不合法. 根据本文算法构造的用于检验用户名合法性的可计数自动机如图 3 所示.

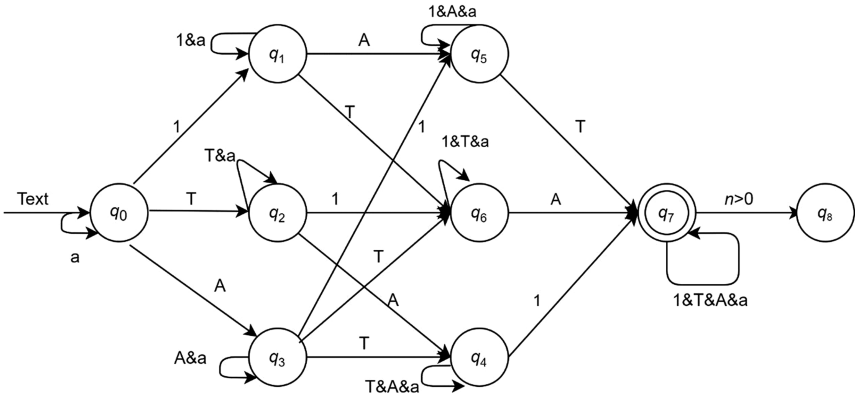

图 3 可计数自动机的状态转移图

本文以输入字符串 s 和 t 为例验证本文构造的自动机的有效性，其中假设 s 为‘abA＊’和 t 为‘2B&’，映射串 $s1$(‘aaAT’)和 $t1$(‘1AT’)由 s 和 t 经映射函数 $f(x)$得到. 图 4 和图 5 分别是 $s1$ 和 $t1$ 的状态转移图，图中的虚线为字符串在自动机中的转移路线. 由图可知，$s1$ 和 $t1$ 经自动机转移到的最终状态分别是 q_4 和 q_8. 由于 q_4 和 q_8 均不是终态，因此这两个字符串是不合法的. 此时解码函数 $h(x)$输出的是：“该用户名缺少数字，不合法”和“该用户名长度不够，不合法”. 上述实例证明本文提出的算法可以有效验证用户名的合法性，且具有简单、稳定的优点.

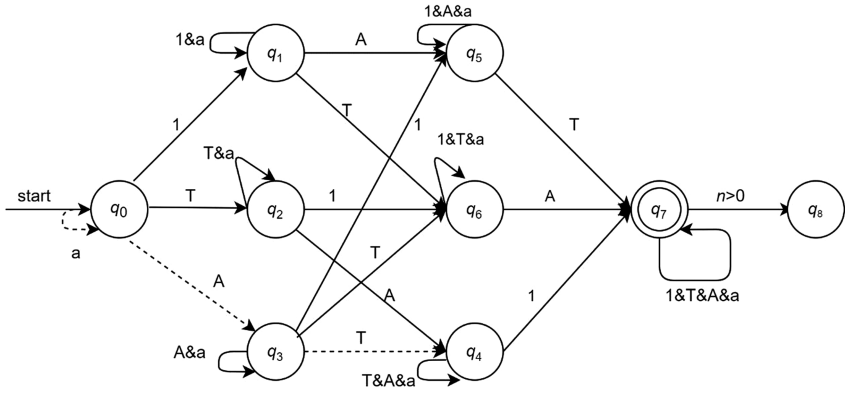

图 4 $s1$ 的状态转移图

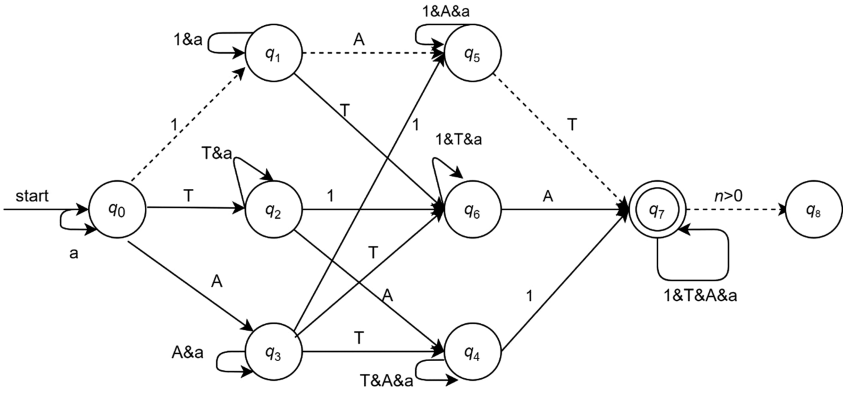

图 5　$t1$ 的状态转移图

3.2　算法复杂度分析

表 3 为传统有限自动机模型(DFA)与本文所提模型(N-DFA)的复杂度，其中 n 为需要满足的条件个数，m 为用户名长度. 由表 3 可知，本文所提模型在时间复杂度和空间复杂度方面均优于传统模型.

表 3　不同模型的复杂度

模型	时间复杂度	空间复杂度
DFA	$O(n)$	$O(2n)+O(m)$
N-DFA	最坏 $O(n)$，最好 $O(1)$	最好 $O(2n)$，最坏 $O(2n)+O(m)$

4　结论

研究表明，本文设计的模型可实现有限状态自动机的计数和设置输入字符串的长度，且具有检测效率高、性能稳定等优点，因此本文方法可用于用户名的合法性检验. 本文算法未能实现计数器和终态的完全分离，拟在今后的研究中通过调整计数器和终态的关系使二者相互独立，以使本文算法达到更好的检测效果.

参考文献：

[1]　王浩，张霖. 基于坏字符序检测的快速模式匹配算法[J]. 计算机应用与软件，2012，29(5)：114-116.
[2]　赵远，秦拯，张大方，等. 一种面向高速网络的模式匹配算法的设计与实现[J]. 微计算机信息，2010，26(12)：167-168.
[3]　付聪，余敦辉，张灵莉. 面向中文敏感词变形体的识别方法研究[J]. 计算机应用研究，2019，36(4)：988-991.
[4]　陈航宇. 正则表达式匹配算法研究[D]. 秦皇岛：燕山大学，2016.
[5]　杨润. 网络数据流的正则表达式匹配技术研究[D]. 哈尔滨：哈尔滨工程大学，2015.
[6]　王安，王芳荣，郭柏苍，等. 基于边缘检测的视差图效果优化[J]. 计算机应用与软件，2019，36(7)：236-241.
[7]　朱宁洪. 字符串匹配算法 Sunday 的改进[J]. 西安科技大学学报，2016，36(1)：111-115.
[8]　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
[9]　曾杰，贲可荣，张献，等. 基于序列模式匹配的 API 误用缺陷检测[J]. 华中科技大学学报(自然科学版)，2021，49(2)：108-114.
[10]　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.
[11]　闫茹，孙永奇，朱卫国，等. 基于 CNN 与有限状态自动机的手写体大写金额识别[J]. 计算机工程，2021，47(9)：304-312.
[12]　姜克鑫，赵亚慧，崔荣一. 基于自动机理论的密码匹配方法[J]. 延边大学学报(自然科学版)，2021，47(2)：141-145.