西文OCR后处理中的有限自动机模型

王 恺 靳简明 王庆人
(南开大学机器智能研究所,天津 300071)
E-mail:wangkai@expervision.com.cn

摘 要 在西文OCR中,从候选结果中挑选最佳结果的后处理操作是必不可少的,并且利用单词拼写检查进行后处理是完全可行的。但是,以往的方法分别在不同程度上具有低可靠性和局限性。为此,该文提出将有限自动机模型应用于西文OCR后处理中,该方法有效地将拼写检查和识别结果信息结合起来,克服了以往方法中存在的低可靠性和局限性,并通过实验验证了该方法的有效性。以识别后处理辅助识别,错误率从0.79%降到0.59%;以识别后处理和系统后处理结合辅助识别,错误率降低到0.55%。

关键词 字符串匹配 有限自动机 光学字符识别 文档图像处理

文章编号 1002-8331-(2004)23-0026-04 **文献标识码** A **中图分类号** TP391.1

The Finite Automaton Model in Western Language OCR Post-processing

Wang Kai Jin Jianming Wang Qingren
(Institute of Machine Intelligence, Nankai University, Tianjin 300071)

Abstract: In western language OCR systems, the post-processing of selecting the best result from some candidates is absolutely necessary. Spell-check can provide reliable information for this task. However, there are some limitations in previous methods in different extents. In this paper, the finite automaton model is applied to the post-processing procedure. It combines the spell-check with the character recognition results. Experiment shows the validity of the method. Using the recognition post-processing, the error rate drops to 0.59% from 0.79%. Using both the recognition and the system post-processing, the error rate drops to 0.55%.

Keywords: string matching, finite automaton, optical character recognition, document image processing

1 引言

光学字符识别(OCR)技术的最初研究始于欧洲,德国人Taushek早在1929年就获得一项有关OCR的专利。为了把大量纸张文档上的文字信息电子化,并进一步利用计算机处理信息,欧美国家自50年代起就开始计算机西文OCR技术研究。成熟的西文OCR商业软件在90年代初陆续出现。

相对于中文汉字、日文汉字以及韩文字符等东方文字而言,西文具有以下特点:

(1)字符结构简单、笔划稀少,字体繁多,仅仅依靠识别器无法完全区分相似字符,往往会得到多个候选识别结果,并且正确结果的可信度不一定是最高的。

(2)以单词作为句子的基本成分,单词之间以空白分割,便于根据词典对单词识别结果进行拼写检查以验证其正确性。

(3)单词中的字符被替换后,成词率低。[1]中对一个字符被替换后的单词的成词率进行了统计:当不区分大小写时平均成词率为0.544%,当区分大小写时成词率只有0.310%。

可见,从候选结果中挑选最佳结果的后处理操作在西文OCR中是必不可少的,并且利用单词拼写检查进行后处理是完全可行的。

当前,已经进行了大量关于拼写检查后处理方面的研究工作。[1]以动态规划的方法计算识别结果与单词的编辑距离,然后以编辑距离最小的单词替换识别结果。这种方法的优点是能够保证单词拼写的正确性,但由于没有充分利用识别信息,有可能会把本来识对的字符改错。[2]将PPM(Prediction by Partial Matching)模型应用于字符切分和识别结果纠正中。字符切分,是将单词中每个字符图像分割出来的一种操作,该操作往往与字符识别和拼写检查迭代进行。在字符切分中,它根据拼写检查结果验证字符切分的有效性,并决定是否需要重新进行字符切分。在识别结果纠正中,它在训练集上对各种错误情况出现的概率进行统计,然后将统计结果用于识别结果纠正中。该方法的优点是利用错误情况的先验信息,在同质量的样张上可以取得较好的效果,但对于不同质量的样张,错误情况发生变化,该方法就会失效。[3]将专家词典用于识别结果验证中,大大节省了人工验证的时间,但其只能作为OCR的一种辅助工具,并不能提高OCR的识别率。

总的看来,以往的方法分别在不同程度上具有下述问题:

(1)低可靠性:对词典的依赖性大,如果识别出来的单词结果在词典中找不到,就有可能被另一个单词替换,导致本来正确的结果被改错。

(2)局限性:没有充分利用识别结果信息,实质上,当前识别器的性能已经很好,虽然对一些相似字符的区分能力还比较差,但正确结果往往可以在候选结果中找到。

基金项目:国家自然科学基金天元基金项目(编号:TY10026002-04-04-01)资助
作者简介:王恺,男,博士研究生,研究方向:文档图像处理,人工智能。

为此，该文提出了一种有限自动机模型，该方法有效地将拼写检查和识别结果信息结合起来，克服了以往方法中存在的低可靠性和局限性。文章组织如下：第二部分首先简要介绍有限自动机理论，然后详细描述应用于西文 OCR 后处理中的有限自动机模型；第三部分讲述如何根据词典构造有限自动机模型；第四部分是系统实现及实验结果；最后，结论在第五部分中给出。

2 有限自动机

2.1 有限自动机理论

定义 1 一个确定的有限自动机(简称 FA)是一个五元组[4]：

$M=(Q,\Sigma,\delta,q_0,F)$

其中：

(1)Q 是有限状态集；

(2)Σ 是有限的输入字符表；

(3)δ 是转移函数，它将 $Q\times\Sigma$ 映射到 Q；

(4)$q_0\in Q$，是初始状态；

(5)$F\subseteq Q$，是终结状态集。

定义 2 一个非确定的有限自动机(NFA)是一个五元组[4]：

$M=(Q,\Sigma,\delta,q_0,F)$

其中 Q，Σ，q_0 和 F 与确定的有限自动机中的含义相同，只是转移函数 δ 不同，它是从 $Q\times\Sigma$ 到 2^Q(Q 的一切子集的集合)上的映射。

2.2 有限自动机模型

目前，有限自动机模型已经被广泛地用于模式匹配问题中。[5]将确定有限自动机模型应用于 BBS 信息监测系统中，以监测 BBS 系统的反动、黄色、敏感信息。[6]将确定有限自动机模型用于防火墙中，以监测数据包。[7]将非确定有限自动机模型应用于 XML 数据过滤系统中，以提供给用户需要的数据，减少用户人工检索的时间。[8]和[9]分别将非确定有限自动机模型应用于蛋白质模式搜索问题和近似串匹配问题中。

西文 OCR 的拼写检查后处理在整个 OCR 系统中主要有两处应用。

(1)识别后处理：在西文文档图像中，字符间粘连的情况普遍出现。因此，字符切分后，往往要利用识别结果及拼写检查对其进行验证，以使字符切分达到最优。把这种用于验证字符切分正确性的拼写检查称为识别后处理。

(2)系统后处理：在所有单词都切分-识别完毕后，在输出结果前，需要对未通过拼写检查的单词进行进一步的处理，以从候选中选出最佳识别结果。把这种用于待输出结果的拼写检查称为系统后处理。

识别后处理和系统后处理主要有以下不同：

(1)根据图 9 可以看出，识别后处理与切分-识别过程迭代进行，直至达到满意的切分-识别结果，而系统后处理比较孤立。因此，对同一个单词，识别后处理会进行一次或多次，而系统后处理最多只进行一次。

(2)识别后处理中的拼写检查要求全匹配，而系统后处理中的拼写检查要求近似匹配。

由于识别后处理和系统后处理的要求不同，因此，两种不同的有限自动机模型分别被应用于识别后处理与系统后处理中，如图 1 和图 2 所示，其中 x 表示空状态，◯表示非终态，◎表示终态。假设词典中有三个字符串：aecdf，aeedf 和 ace。

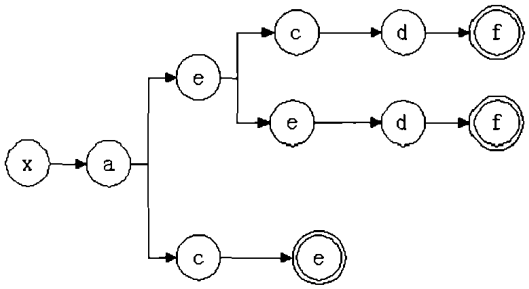

图 1 识别后处理中的有限自动机模型

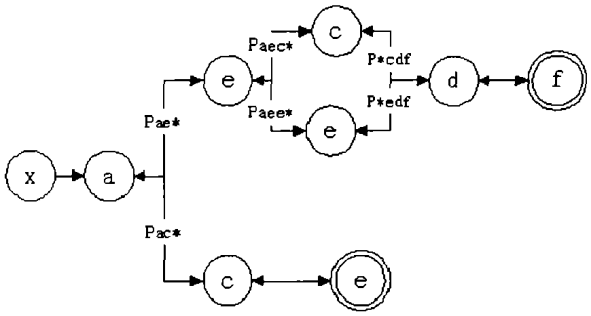

图 2 系统后处理中的有限自动机模型

由于识别后处理中的拼写检查要求全匹配，因此其有限自动机模型比较简单，只需将单词中的字符按从左到右的顺序输入，最后能够被终态接收的单词即为合法单词。通过识别后处理，可以将正确结果从候选结果中挑出来。但是，由于该有限自动机模型具有以下局限性：

(1)无法处理包含拒识字符的单词。如 aecdf 被识别成 $(a,e)(a,e)\sim(b,d)(f)$，其中~表示拒识，$(X1,X2,\ldots,Xn)$ 表示结果可能是 $X1,X2,\ldots,Xn$ 中的任一个，并且 $P(X1)<P(X2)<\ldots<P(Xn)$($P(X)$表示识别可信度)。如果不进行拼写检查，那么最终的识别结果就会是 aa~bf。

(2)无法处理具有二义性的识别结果。如识别结果为 $(a,e)(a,e)(a,c,e)(b,d)(f)$，那么能够通过拼写检查的单词有两个：aecdf 和 aeedf。

为了弥补识别后处理的局限性，系统后处理使用图 2 所示的有限自动机模型，双向弧上的标识表示该分支的概率，如 $Paec*$ 表示字符串 ae 后面接 c 的概率，$P*cdf$ 表示字符串 df 前面接 c 的概率。显然，对于图 2 所示的有限自动机模型来说，$Pae*+Pac*=1$。系统后处理中有限自动机模型的运作过程如下：

(1)将识别结果从有限自动机的头部输入；

(2)当遇到拒识字符时，则将当前分支截取下来生成一个新的有限自动机，并在该分支的尾部加一空状态 x，作为新生成的有限自动机的头部，将识别结果倒置输入到新的有限自动机的头部。如 aecdf 被识别成 $(a,e)(a,e)\sim(b,d)(f)$，将结果输入到图 2 所示的有限自动机的头部，在遇到拒识字符时，识别结果变为 $ae\sim(b,d)(f)$，并到达如图 3 所示的分支，将该分支截取下来生成的新的有限自动机如图 4 所示。将识别结果倒置变为 $(f)(b,d)\sim ea$，输入到图 4 所示的有限自动机的头部，最后可以确定识别结果为 $ae\sim df$；

(3)当遇到具有二义性的识别结果时，则优先选择具有高概率的分支进行。如识别结果为 $(a,e)(a,e)(a,c,e)(b,d)(f)$，并且 $Paec*>Paee*$，则最终确定的识别结果为 aecdf。

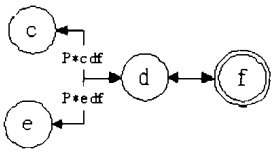

图3 有限自动机分支

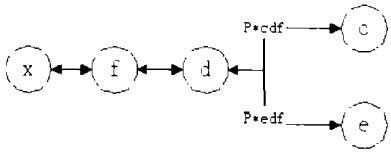

图4 分支生成的有限自动机

这样，通过识别后处理和系统后处理，就可以将识别结果从候选结果中正确地挑选出来。

3 有限自动机模型的构造

识别后处理中的有限自动机模型构造过程如下：

(1)建立一有限自动机，使其仅包含空状态 x；

(2)从词典中顺序取出单词，若有单词可取，则将单词送入有限自动机中，否则构造完毕退出；

(3)当运行到状态 y，且无适合分支可选择时，则将剩余 n 个字符作为 n 个状态接在状态 y 的后面，转(2)。

假设词典中有5个字符串，依次为 abcde，abdde，bcefg，bd，bdfgk，则识别后处理中的有限自动机模型构造过程如图5所示。

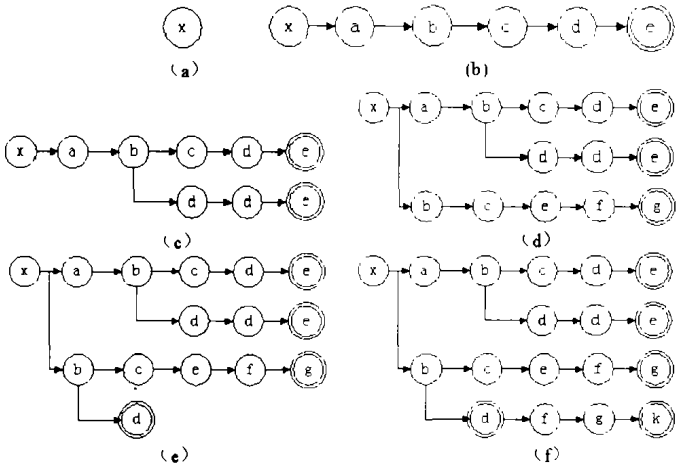

图5 识别后处理有限自动机模型构造过程

为了生成系统后处理中的有限自动机模型，文章将构造过程分为两部分：前向学习和反向学习。其中，前向学习与识别后处理有限自动机模型构造过程类似，只是分支弧具有计数功能(初值为0)，分支被选择一次，则对应弧上的数值加 n(n 为待学习单词出现的次数)。

仍然假设词典中有5个字符串，依次为 abcde(5)，abdde(3)，bcefg(4)，bd(10)，bdfgk(8)，其中括号中的数字表示该字符串的出现次数，则前向学习过程如图6所示。

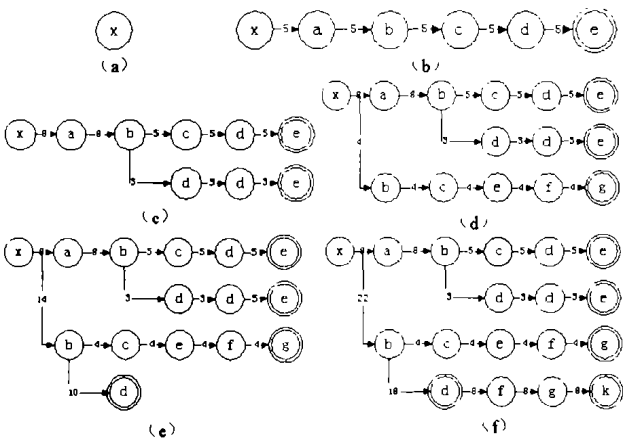

图6 前向学习过程

定义3 终态节点所在的层称为第1层，指向第1层节点的节点所在的层称为第2层，依次可以定义第3层，第4层……。如图6(f)中，第1层包括 e,e,g,k,d 五个节点，第2层包括 d,d,f,g,b 五个节点。

定义4 节点 o 的下一个节点 q 到终态的路径称为节点 o 的一个分支。如图6(f)中，位于 a 后面的节点 b 包含两个分支：cde 和 dde。

在前向学习的基础上，反向学习过程如下：

(1)令当前层 n=3；

(2)如果第 n 层节点数为0，反向学习过程结束退出，否则转(3)；

(3)查找第 n 层节点的分支，如果某节点具有两条或两条以上的分支数，则将该节点的分支从尾部开始对具有相同状态值的节点进行合并操作；

(4)n=n+1，转(2)。

对图6(f)的反向学习过程如图7所示。

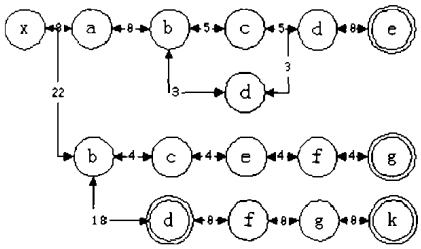

图7 反向学习过程

最后，根据前向分支计算 A^* 的概率，根据反向分支计算 $^*A$ 的概率(A 表示字符串)，则最终的系统后处理有限状态机模型如图8所示(分支缺省概率为1)。

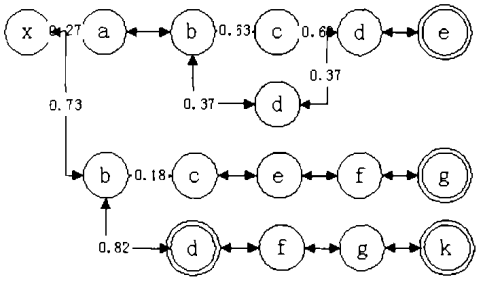

图8 系统后处理有限状态机模型

4 系统实现及实验结果

从大量网页上采集单词，并对每个单词的出现次数进行统计，然后根据第3部分中介绍的构造方法分别生成识别后处理有限自动机和系统后处理有限自动机。以南开大学开发的RTK6.0为OCR引擎，通过图9所示的系统控制进行实验。实验分为3部分：纯识别的系统I、仅有识别后处理辅助的系统II、识别后处理和系统后处理共同辅助的系统III。实验结果如表1所示。

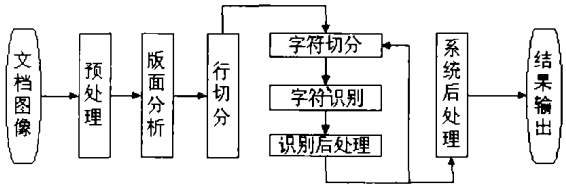

图9 系统流程控制

表 1 系统识别率

总字符数	系统 I		系统 II		系统 III	
	错误数	错误率	错误数	错误率	错误数	错误率
4116640	32474	0.79%	24367	0.59%	22654	0.55%

由表 1 可以看出，通过识别后处理和系统后处理，错误率降低了近 1/3，达到了预期的效果。

5 结论

该文将有限自动机模型应用于西文 OCR 后处理中，该方法有效地将拼写检查和识别结果信息结合起来，克服了以往方法中存在的低可靠性和局限性，并通过实验验证了该方法的有效性。以识别后处理辅助识别，错误率从 0.79%降到 0.59%；以识别后处理和系统后处理结合辅助识别，错误率降低到 0.55%。（收稿日期：2004 年 4 月）

参考文献

1.吕学强，迟呈英.英文光学字符识别的后处理[J].鞍山钢铁学院学报，2002;25(3):192~196

2.W J Teahan,S Inglis,J G Cleary et al.Correcting English text using PPM models[C].In:Data Compression Conference,1998:289~298

3.Hauser SE,Browne AC,Thoma GR et al.Lexicon assistance reduces manual verification of OCR output[C].In:Proc 11th IEEE Symposium on Computer-Based Medical Systems,1998:90~95

4.陈有祺.形式语言与自动机[M].南开大学出版社，1999

5.罗光春，李炯.有限自动机在 BBS 信息监测系统中的运用[J].电子科技大学学报，2002;31(3):262~265

6.James Moscola,John Lockwood,Ronald P Loui et al.Implementation of a Content-Scanning Module for an Internet Firewall[C].In:11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines,2003:31~38

7.Yanlei Diao,Michael J Franklin.High-Performance XML Filtering:An Overview of YFilter[J].IEEE Data Engineering Bulletin,2003;26(1):41~48

8.Gonzalo Navarro,Mathieu Raffinot.Fast and simple character classes and bounded gaps pattern matching,with applications to protein searching[C].In:Proceedings of the 5th Annual International Conference on Computational Molecular Biology,2001:231~240

9.Heikki Hyyro,Gonzalo Navarro.Faster Bit-parallel Approximate String Matching[C].In:Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching,2002:203~224

（上接 5 页）

$$L_y=L\times\cos(\gamma) \tag{8}$$

枝条挠曲方程$\bar{y}_L$所表示的曲线即是枝条趋光作用下的形态结构，如图 3(b)。枝条在自重与趋光作用下总的挠度$\bar{y}$、端截面转角 $\varphi(t)$分别为两者的线性迭加：

$$\bar{y}=\bar{y}_w+\bar{y}_L=-\frac{L_y\bar{x}^2}{6El}(3l-\bar{x})-\frac{w_y\bar{x}^2}{24El}(\bar{x}^2-4l\bar{x}+6l^2) \tag{9}$$

$$\varphi(t)=\varphi_w(t)+\varphi_L(t)=-\frac{L_y l^2}{2El}-\frac{w_y l^3}{6El} \tag{10}$$

挠曲方程(9)反映枝条在自重与趋光作用下的变形情况，同时也是该时刻枝条的形态，它说明枝条变形与反映枝条柔软度的杨氏弹性模量、惯性矩、枝条自重、枝条趋光的强弱以及枝条的长度有关。此外，端截面转角也反映了枝条在自重与趋光作用下的变形情况，$\varphi(t)$越大，变形越大。

新旧坐标系关系：

$$\begin{pmatrix}x\\y\end{pmatrix}=T\begin{pmatrix}\bar{x}\\\bar{y}\end{pmatrix}, \tag{11}$$

$$T=\begin{pmatrix}\cos(90-\gamma) & -\sin(90-\gamma)\\ \sin(90-\gamma) & \cos(90-\gamma)\end{pmatrix} \tag{12}$$

则枝条在(xoy)下的挠曲方程为：

$$x=\bar{x}\times\sin(\gamma)-\bar{y}\times\cos(\gamma) \tag{13}$$

$$y=\bar{x}\times\cos(\gamma)+\bar{y}\times\sin(\gamma) \tag{14}$$

将式(9)与式(13)、(14)联立，则式(13)、(14)是一个以$\bar{x}$为参数的曲线方程。而$\bar{x}$是时间 t 的函数，因此，式(13)、(14)是一个时间的复合函数，即枝条的形态是一个动态的结构。

光垂直向下照射，等效为一个垂直向上的拉力，如图 3(c)。在坐标系$(\bar{x}0\bar{y})$下，挠曲方程与端截面转角为：

$$\bar{y}=\varepsilon(t)=\frac{L_y\bar{x}^2}{6El}(3l-\bar{x})-\frac{w_y\bar{x}^2}{24El}(\bar{x}^2-4l\bar{x}+6l^2) \tag{15}$$

$$\varphi(t)=\frac{L_y l^2}{2El}-\frac{w_y l^3}{6El} \tag{16}$$

$$L_y=L\times\sin(\gamma),w_y=w\times\sin(\gamma) \tag{17}$$

式(11)、(12)代入式(15)，得$(x0y)$坐标系下的变形的挠曲方程。枝条生长与形态生成过程如图 4。

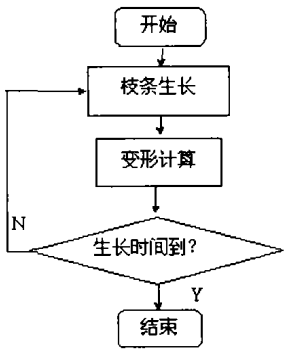

图 4 枝条生长与形态生成过程

4 结束语

该文将植物的生长视为三维状态空间中状态矢量的运动过程，将分生组织的作用作为系统的输入量，将植物枝条视为状态矢量的运动轨迹；同时，综合考虑了重力与趋光性对枝条形态的变形作用，增加了植物形态的真实性。该模型在虚拟植物形态结构生成方法上，以及寻找建立植物形态结构与生长机理的关系上做了有益探索，为虚拟农林以及计算机动画等提供了具有实际应用价值的研究方法。（收稿日期：2004 年 4 月）

参考文献

1.郝小琴.森林景物的三维迭代函数系统建模技术的研究[J].计算机学报，1999;22(7):768~773

2.Prusinkiewicz P,Hammel M,Mjolsness E et al.Animation of Plant Development[C].In:Proceedings of SIGGRAPH,1993:351~360

3.赵星，熊范纶，Philipped Reffye.一种新的植物枝条弯曲生成算法[J].中国科学技术大学学报，2001;31(6):714~720

4.明延凯.植物学教程[M].第一版，北京：石油大学出版社，1991

5.刘豹.现代控制理论[M].第二版，北京：机械工业出版社，2001

6.刘鸿文.材料力学(上)[M].第二版，北京：高等教育出版社，1982