doi: 10.3969/j.issn.1006-1576.2010.09.028

基于 OCR 的拼写校正系统

赵莉
（西安工业大学 计算机科学与工程学院，陕西 西安 710032）

摘要：针对光学字符识别（Optical Character Recognition，OCR）过程中出现的英文字母识别错误问题，通过分析其出错原因，提出一种将拼写错误、OCR 错误规则和编辑距离法相结合的校正算法，并实现了拼写校正系统最重要的 2 项功能：拼写检查和拼写校正。其中，拼写检查模块采用了查字典技术；校正过程中则采用了编辑距离法。通过选取 5 种常用字体的打印档的辨识情况，测试了算法的有效性。实例证明，校正后的识别率都提高了 2%～4%。

关键词：拼写校正；光学字符识别；编辑距离；OCR 距离
中图分类号: TP311; TP301.6　　**文献标识码**：A

Spelling Correction System Based on OCR

Zhao Li
(Dept. of Computer Science & Engineering, Xi'an University of Technology, Xi'an 710032, China)

Abstract: Through the analysis of English spelling errors of OCR (optical character recognition), introduce an algorithm of combination of phonetic spelling correction, OCR regulations and edit-distance method. And spelling check and spelling correction are two most important functions of the spelling correction system. In which, dictionary database verification technology is applied in the spelling check module and the edit-distance method is applied in correction process. Chooses five common printed character types to print and test the effect of the algorithm. The test shows that the recognition rate improves 2-4 percents.

Keywords: phonetic correction; OCR; edit distance; OCR distance

0　引言

光学字符识别（Optical Character Recognition，OCR）简称文字识别，是文字自动输入的一种方法。它通过扫描和摄像等光学输入方式获取纸张上的文字图像信息，利用各种模式识别算法对文字形态特征进行分析，判断出文字的标准编码，并按通用格式存储在文本文件中，是一种非常快捷、省力的文字输入方式。

拼写校正可以帮助输入者发现拼写错误的单词，并对错误单词进行校正，将校正后的结果提供给输入者使用。当 OCR 对一个英文单词扫描完成后，会给出一个待识别的候选词。如果该候选词是可识别的，则为识别结果；如果识别错误，则给出一个最接近词。故针对 OCR 扫描结果的特点，提出一种英文拼写校正算法。

1　概述

一个拼写校正系统必须能准确地检测出输入词的拼写错误，与用户预期的正确单词相比，这些错误可能是缺字母、多字母、错字母等，如表 1。

故定义如下：

1) 非词错误：例如，将 correct 拼成 corect，correct 是一个英文单词，而 corect 不是一个英文单词；2) 真词错误：例如，将 week 拼成了 weak，但 week 和 weak 均为一个有完整意义的英文单词。

笔者只针对英文文本中的非词错误编写算法，不涉及真词错误。

表 1　拼写错误的单词和拼写正确的相近单词

拼写错误的单词	拼写相近的正确单词	错误原因
inpat	input	字母 u 有错
corect	correct	缺字母 r
worde	word	多字母 e
frend	friend	缺字母 i
	frond	字母 o 有错

2　算法说明

拼写校正系统最重要的功能是拼写检查和拼写校正。

为了快速、准确地识别出英文文本中拼写错误的单词，拼写检查模块采用错误检测技术。识别方法主要包括 N-gram 分析法和查字典法 2 种。

N-gram 分析法的主要思想是：对于输入串中的每一个 n 元串（n 一般取 2 或 3），在事先编辑好的一个 N-gram 表中进行查找，看它是否在表中存在以及在表中出现的频次，那些不存在或出现频次非

收稿日期：2010-04-14；修回日期：2010-05-05
作者简介：赵莉（1972-），女，陕西人，副教授，从事软件工程、数据仓库、数据挖掘技术研究。

常低的 n 元串被认为是可能的拼写错误，如“ZZK”就是错误的三元串。N-gram 分析法通常需要一个字典或大规模的文本语料，以便事先编辑 N-gram 表。

由于基于查字典法的校对系统查错精度高，故该采用查字典法。其主要思想是：检查所输入的 n 元串是否在字典或可接受的词表中，如果在字典中，则是正确的单词；否则，认为该输入串是一个拼写错误的单词。查字典法所用字典必须选择合适的词汇量。如果词汇量太小，一些专有名词、人名或地名等生僻词会被认为是错误单词。拼写检查所使用的字典需要针对不同的专业领域生成各自的字典，并且字典可以自学习，能随时引入新词到字典中。

为了将拼写错误的单词识别出来，找出这个错词最可能的正确拼写方法，需要了解拼写错误是如何产生的。例如，在光学扫描识别时，识别程序通过扫描所获得的图片来识别单词，较常产生的错误如下：i→l，如 life 识别成 llfe；rn→m，如 return 识别成 retum。而这些错误在用户使用键盘输入时绝少发生，必须根据应用的类型设计有针对性的拼写校正算法。

目前，最常用的字符串相似度计算方法是编辑距离法（Edit Distance），编辑距离又称为来文史特距离（Levenshtein Distance），定义如下：

令字符串 $P=p_1p_2\cdots p_n$ 和 $W=w_1w_2\cdots w_m$ 是有穷字母表 Σ 上的 2 个字符串，ε 表示空字符串。编辑操作是一个二元组（a,b），也可表示为 a→b，其中 $a,b\in\Sigma\cup\{\varepsilon\}$，$(a,b)\neq(\varepsilon,\varepsilon)$，$a\neq b$。令 W 由 P 通过编辑操作 a→b 生成，P=xay，W=xby，x 和 y 为字符串。若 $a\neq\varepsilon$，$b\neq\varepsilon$，则 a→b 称为替换操作；若 $a=\varepsilon$，$b\neq\varepsilon$,则 a→b 称为插入操作；若 $a\neq\varepsilon$，$b=\varepsilon$，则 a→b 称为删除操作。字符串 P 和 W 的编辑距离 ed(P,W) 是由 P 变换到 W 所需要的最少编辑操作数目。计算 2 个字符串编辑距离的标准算法是 Wagner 和 Fisher 提出的动态规划算法。令 P 和 W 的长度分别为 n 和 m，计算 P 和 W 的编辑距离 ed(P,W)的过程就是给一个 n 行、m 列的矩阵 edit 赋值的过程。edit 矩阵赋值过程如下：

edit(0,0)=0
edit(i,0)=i
edit(0,j)=j
edit(i,j)=min(edit(i-1,j)+1, edit(i,j-1)+1,
　　edit(i-1,j-1)+ed(P_i,W_j))

其中，Pi 和 Wj 分别表示 P 的第 i 个字符和 W 的第 j 个字符。若 Pi=Wj，则 ed(Pi,Wj)=0，否则 ed(Pi,Wj)=1。该算法的时间复杂度和空间复杂度均为 O(n×m)，算法执行结束后，edit 矩阵第 n 行第 m 列元素值即为 P 与 W 的编辑距离。

编辑距离法就是将一个单词经过增、删、改的操作转化为另一个单词所要的最少操作次数，编辑距离不考虑“输入方式”产生错误的可能性大小，而将每次改变的权值定为 1。例如，错误单词 wcld 和 wold 的编辑距离为 1，wcld 和 wild 的编辑距离为 1，从 OCR 识别特点的角度，字母“o”和“c”之间的错误可能性要高于“o”和“I”之间的错误可能性，因而选择校正结果时，wold 是更接近的词。可以看出，这些拼写错误是 OCR 识别系统固有特点所产生的，观察和总结这些出错的原因，并提供有针对性的校正算法，可更准确地解决 OCR 系统中的拼写校正问题。

在校正过程中，OCR 识别较易产生的拼写错误和正确词之间的距离，称之为 OCR-Distance，它是对编辑距离的修正，并且应当比编辑距离小。下面比较 OCR-Distance 和 Edit-Distance：OCR-Distance 是 OCR 识别过程中较易产生的拼写错误，Edit-Distance 是不考虑任何因素的 2 个单词之间的编辑距离。

在满足 OCR 替换规则的条件下，2 个词之间的 OCR-Distance 和 Edit-Distance 满足 OCR-Distance< Edit-Distance。例如 room→roorn，依照 OCR 识别的结果，m 和 rn 是一个 OCR 替换规则，此处这 2 个词之间的 OCR-Distance <Edit-Distance；

OCR 识别时，存在一种情形：可以识别出某个位置有字母，而不确定是何字母。笔者把这个识别出来的字母用~号替代。例如，Wear 和 ~ear 的编辑距离为 1，ear 和~ear 的编辑距离也为 1，而 wear 有 4 个字母，更接近~ear，因此 ear 和~ear 的 OCR-Distance 要大于 Wear 和~ear 的 OCR-Distance。

受限于识别系统词库的大小，不能识别出一些专有名词。故对于大于一定 OCR-Distance 的单词，应当不予校正。

参照编辑距离的定义，提出 OCR-Distance 的定义，用于计算 2 个字符串的 OCR-Distance。OCR-Distance 定义如下：

令 P 和 W 的长度分别为 n 和 m，计算 P 和 W 的 OCR-Distance od(P,W)的过程就是给一个 n 行、m 列的矩阵 OCRD 赋值的过程。OCRD 矩阵赋值过程如下：

OCRD (0,0)=0
OCRD (i,0)=i * 2
OCRD (0,j)=j * 2
OCRD (i,j)=min(OCRD (i-1,j)+2, OCRD (i,j-1)+2,
OCRD (i-1,j-1)+od(P_i,W_j))

而对于 od(Pi,Wj)函数，除了考虑 Pi、Wj 是否相等外，加入如下判断规则：

如果 Pi= ～ 并且 Pi−1=Wi−1，Pi+1=Wi+1 则 od(Pi, Wj)为 1.5；

如果 (Pi,Pi+1…) 和 (Wj,Wj+1…) 满足预先定义的 OCR 替换规则，则替换位置处的 od(Pi,Wj) 为 1；

如果被替换的是多个字母，则被替换的每一个 od(Pi+k,Wj+l) 值都为 0；

执行完成后，OCRD(n, m)就是 OCR−Distance。

注意：这里所提的 OCR 替换规则，是基于光学扫描识别易出错误的替换规则。部分规则如表 2。

表 2 OCR 规则替换表

原串	可换串	原串	可换串
i	l	vv	w
c	o	v	y
rn	m	q	p
b	d	o	d
f	t	n	h
…	…		

对校正比对范围内的每一个词，计算 OCR−Distance，并按照从小到大的次序排列，最终排在队列最前面的词就是选中的候选词。

例如，在新的算法里，错误的输入词 flll 的候选词和 OCR−Distance 如表 3。

表 3 flll 的 OCR−Distance

候选词	OCR−Distance	说明
fill	1	l 和 i 做 OCR 替换
fall	2	l 和 a 做编辑替换
fail	3	a 和 l 编辑替换， i 和 l OCR 替换
foil	3	o 和 l 编辑替换， i 和 l OCR 替换
fille	3	i 和 l 做 OCR 替换，删掉字母 e

在这种校正方法下，候选词 fill 超过了 fall，排到了最前面，即 fill 是 flll 的 OCR 最接近词。

拼写校正程序除了拼写检查和识别功能外，还需要采用一个的合适的字典，最好是有针对性的字典。字典太小，很多专业性的词汇、人名、地名可能会被认为是错误的单词；而字典太大，会占用更大的存储空间，增加匹配时间，降低准确率。针对 OCR 系统的英文拼写校正程序，在拼写检查时，应使用较大的字典，这样识别出的候选单词就较多，降低了拼写错误的可能性；而拼写校正时，采用较小的字典，这样校正结果集较小，也更准确。

3 结果分析

为了测试设计的拼写校正算法的有效性，依照不同的条件设计了测试案例。选取 5 种常用字体的打印档的辨识情况，字号均为 12 号，通过喷墨打印机输出到纸质文档，环境光照度 50～55LX，采用某手持式测试平台，测试总字符数 10 232 个，测试结果如表 4。

表 4 测试结果

字型	单词内间距 (pixels)	单词间间距 (pixels)	行距 (pixels)	OCR 识别率	校正后识别率
Arial	1－3	14－19	26－42	96.7%	98.7%
Courier New	3－9	26－28	26－44	85.7%	88.3%
Garamond	1－2	11－13	24－48	90.9%	94.8%
Bookman Old Style	1－4	15－18	27－48	97.3%	98.7%
Times New Roman	1－2	7－14	26－45	94.5%	96.1%

由测试结果可以看到，对于各种不同字体所产生的结果，校正后的识别率都提高了 2%～4%，从而验证了该算法的有效性。

4 结论

该算法已应用到实际系统中，获得了较好的应用效果。关于 OCR 的拼写校正相关的算法和方向很多，更深层次的研究可以把错误发生的上下文也考虑进去，通过语法检查拼写错误。例如“I come form Beijing”，如果单纯检测单词是不能发现问题的，但是通过上下文关联，就会发现 form 应为 from，这是下一步继续深入研究的课题。

参考文献：

[1] Harriet Wittels，Joan Greisman. The Clear and Simple How to Spell It: A Handbook of Commonly Misspelled Words［M］. Americian:Grosset & Dunlap, 2007: 5.

[2] 王永生, 李敏. 英文文语转换系统中基于形态规则和机器学习的重音标注算法[J]. 计算机应用, 2008, 28(1): 88−91.

[3] 樊娜, 蔡皖东, 等. 中文文本情感主题句分析与提取研究[J]. 计算机应用, 2009, 29(4): 1171−1173.

[4] 李晓光, 王鹏, 等. 面向多领域资源的汉英双语语料库构建的研究[J]. 计算机应用, 2008, 28(1): 146−148.

[5] 孙巍. 一种面向中文信息检索的汉语自动分词方法[J]. 现代图书情报技术, 2006, 1(7): 33−36.

[6] 王恺新. 西文 OCR 后处理中的有限自动机模型[J]. 计算机工程与应用, 2004: 23.

[7] 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, 25(3).

[8] 张仰森, 俞士汶. 文本自动校对技术研究综述[J]. 计算机应用研究, 2006(6).

[9] 龚才春, 黄玉兰, 许洪波. 基于多重索引模型的大规模词典近似匹配算法[C]. 第三届全国信息检索与内容安全学术会议, 2007: 7

[10] 马金山, 刘挺, 李生. 基于 n-gram 及依存分析的中文自动查错方法[C]. Advances in Computation of Oriental Languages--Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, 2003.