无字库智能造字中汉字基元的统计分析与预测

鄢 琦,骆仁波,皮佑国
(华南理工大学 自动化科学与工程学院,广东 广州 510641)

摘 要:针对建立长期稳定和规模合理的字库标准这一难题,提出了基于认知机理的无字库智能造字系统,以汉字基元库代替汉字字库。文中在介绍无字库智能造字机理及其汉字基元的基础上,就该汉字基元库进行基元统计分析与预测,运用回归分析的方法,得到拟合模型方程,运用该方程拟合出基元库中基元数量随着汉字数量增加的变化规律曲线,从而预测出10万汉字的时候基元的数量,证明了在已有的全部10万汉字的情况下,基元数量在现有基础上增加不多。

关键词:智能造字;汉字基元;回归分析;曲线拟合
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2012)04-0033-04

Statistical Analysis and Prediction of Prototype on Chinese Character Intelligent Formation System Without Font Character

YAN Qi, LUO Ren-bo, PI You-guo
(College of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: Aimed at the problem of forming an appropriate scale and long-term stability of the Chinese character library standards, the Chinese character intelligent formation system without font character based on cognitive mechanism is proposed, replacing the font character library with the Chinese character prototype library. Based on the mechanism of Chinese character intelligent formation system, conduct a statistical analysis and prediction for the Chinese character prototype library and use the regression analysis method to get the model equation which the prototype number changes with the Chinese characters increasing, finally predict the prototype amount of one hundred thousand Chinese characters to prove that the number of prototypes increase not much on the existing amount with all one hundred thousand Chinese characters.

Key words: intelligent making chinese character; chinese character prototype; regression analysis; curve fitting

0 引 言

中文汉字信息化[1]发展到目前,已达较高水平,中文信息化技术的研究开发在输入技术、模式识别和自然语言理解等领域也取得了不少成果,为我国的信息化做出了不可磨灭的贡献。

截至目前,我国和国际组织开发的中文信息系统都采用汉字字库作为底层核心技术,以汉字作为信息处理的最小单位。这种方式先将某一标准规定的汉字建立字库,并对每个汉字进行唯一的地址编码。该地址编码作为计算机内部存储、传输和管理等信息处理的对象,输入时根据交换码到字库中选字。此种方式虽基本满足目前我国社会对信息化的需求,但也存在着难以建立规模适度和长期稳定的汉字信息化标准等固有的弊端[2]。

鉴于汉字字库的不足,不少学者对汉字生成技术进行了研究和探索,并取得了一些研究成果。上世纪九十年代有学者提出智能汉字库模型[3,4]。文献[5]针对汉字信息处理中出现的缺字问题,提出采用组件拼合的方法来造字。文献[6]则针对汉字字库存储量大的问题,提出了一种基于部件复用的分级汉字字库构想,将汉字中重复使用的部件和一些基本笔画归纳总结,形成标准部件库。但上述文献仍是基于汉字字库,着眼于解决汉字生僻字的输入问题。文献[7]提出了智能造字的结构框架,用汉字基元库取代汉字字库。那样,汉字基元就相当于拼音文字的字母,不管汉字如何发展,汉字基元是长期稳定不变的。汉字是由象形、指事字及其符号按照会意、形声法则拼合而成的文字系统,象形、指事字及其符号就是汉字的基本元素——基元。汉字基元可在辞书的部首基础上通过实验获得。如果汉字基元能通过已有汉字实验并有较合理的统计分布则可以作为标准确定下来,新发展的汉字按照基元标准组成,则汉字就可以像拼音文字一样,基元标准长期稳定不变。相关文献[8~11]在此基础上

收稿日期:2011-09-07;**修回日期**:2011-12-11
作者简介:鄢 琦(1986-),女,硕士研究生,研究方向为模式识别与智能信息处理;皮佑国,教授,研究方向为模式识别。

进行了大量的实验研究，遴选出了1086个汉字基元，成功造出了70244个能收集到的汉字。

但据有关记载，中国汉字有10万之多，其准确数量尚不得而知，更别说能够收集到并进行实验。要完成8～10万字的收集在短期内也是不可能的。那么，遴选的1086个汉字基元在多大程度上代表汉字的基本元素，以10万汉字计，还需要增加多少汉字基元，增加定的基元数目多不多，是智能造字必须回答的问题。

文中的思路是：用遴选的汉字基元重复汉字实验，对汉字基元进行统计和分析，用统计数据来进行回归分析，获得汉字数量与汉字基元的模型方程，利用该模型方程预测和估计在10万汉字时汉字基元的数量。

1 无字库智能造字机理及汉字基元简介

1.1 无字库智能造字机理

无字库智能造字是一个将基元库中的基元按照一定的汉字结构和知识规则构造成汉字的过程。其实现原理如图1所示。从图1可以看出，无字库智能造字的实现，要以汉字基元库和汉字结构知识库作为依托。在此基础上，以汉字基元作为汉字造字的基本单元，在智能造字过程中，根据外界汉字编码的输入，汉字基元依据一定汉字结构组合成汉字。

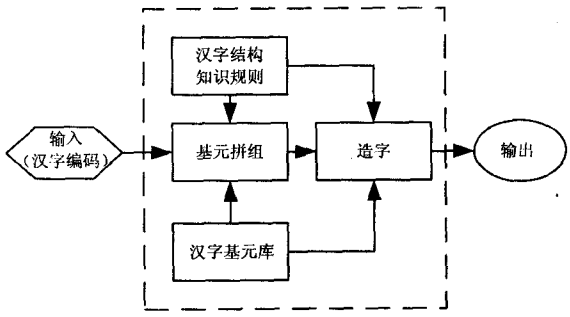

图1 无字库智能造字系统实现原理

1.2 汉字基元简介

由上一节的造字过程可以知道，所讨论的汉字基元就是无字库智能造字系统中用来构成汉字的基本元素，相当于拼音文字中的字母。组成汉字时根据需要调出相应的基元进行组合。无字库智能造字理论认为，如果确定的基元库中的基元能满足已有汉字的需要，那么将该基元集合作为标准固定下来，新出现的字就按照标准规定的基元来构造。这样，确定的基元集合就可以像拼音字母一样长期不变，从而解决了汉字信息化标准长期稳定的问题。由于汉字基元是由象形字和指事字及其符号组成[12]，反映汉字的“义”，能够体现汉字的表意文字特征，从而能够传承汉字文化。汉字基元的提取以辞书中的部首作为初始集合，通过实验方法进行补充完善，再根据统计分布进行调整，以保证提出的基元集合组字时具有较高的效率[11,12]。

基元提取分别对字符集GB2312规定的6763个汉字，字符集GB18030-2000规定的27533个汉字以及字符集GB18030-2005规定的70244个汉字进行实验，三个集合提取的基元数量分别为557个、721个和1086个[1,10,11]。汉字从6763个增加到27533个时增加了164个基元，从27533增加到70244个汉字时，基元增加了361个。部分基元列举如表1所示。

表1 基元列举表

丶	一	丨	丿	乚	冂	亅	乙	𠂇
𠂆	又	亻	乃	亏	入	𠂉	卜	匚
匕	几	夂	宀	讠	儿	八	阝	扌
一	了	冫	𠂉	乚	纟	忄	寸	○
丁	厷	刀	卜	刂	口	口	凵	巛
勹	阝	巳	二	十	巛	宀	犭	九
七	厶	也	子	又	乂	巾	已	刂

但是，现在1086个汉字基元是在7万多汉字的基础上遴选出来的，以10万汉字计，其多大程度反映汉字基本元素，还需加多少汉字基元，是需继续探讨的。

2 造字实验与基元统计分析与预测研究

实验目的是获得汉字造字中汉字数量与使用汉字基元数量之间的关系，从而预测出10万汉字时的基元数。实验方法是对GB18030-2005规定的70244个汉字集合进行造字，造字过程中对基元进行统计，然后运用回归分析得到统计规律模型，用模型曲线来拟合基元的统计离散点，通过计算拟合系数验证模型的拟合精度，借此外推，预测得到结果。

2.1 无字库智能造字实验数据统计

无字库智能造字系统中对上一节中遴选出来的汉字基元在三大字符集中的使用情况进行了统计，统计数据列举如表2所示。

对表2的数据进行反复分析对比得出：

1）使用频度高的基元主要是从常用字符集GB2312-80和GB18030-2000提取出的基元群体，如：“日”。

2）新增的字符一般都是生僻字，这些生僻字包含的基元个数较多，如：“犇、麤”，但是构成它们的基元还都是些频度较高的基元；为了拼合生僻字而新增加的基元（在GB2312-80和GB18030-2000中没有出现，而在GB18030-2005中出现了的基元）使用频度非常低，如表中的“㐱”。

为让字符集GB18030-2005中统计出的基元库数据更准确地反映基元统计规律，避免或减少字符集中汉字顺序的人为排列影响，依据表2分析得到的两点结论，将GB18030-2005中7万多汉字按照单个字符

所含基元数量从大到小进行顺序混合。

在无字库智能造字平台上对基元进行统计得到表3的汉字数和基元数目表。

表2　各基元在三大字符集中使用情况列举

基元	扌	氵	灬	讠	木	力
GB2312-80	289	68	0	152	17	101
GB18030-2000	905	331	0	168	56	410
GB18030-2005	2106	915	1	170	138	988
基元	冊	氏	虫	隶	刖	巛
GB2312-80	19	14	2	0	2	16
GB18030-2000	88	50	77	0	3	78
GB18030-2005	245	88	299	1	31	515
基元	弋	平	大	亠	止	日
GB2312-80	9	1	75	167	22	428
GB18030-2000	37	1	277	628	102	1907
GB18030-2005	106	2	756	1807	327	5156
基元	几	仪	甫	宀	犭	乂
GB2312-80	2	40	0	199	0	70
GB18030-2000	9	185	0	704	0	321
GB18030-2005	30	447	4	2420	1	1174

表3　汉字个数和汉字基元数目表样本集

汉字数	500	1000	1500	2000	2500	…
基元数	483	650	668	688	699	…
汉字数	17500	18000	18500	19000	19500	…
基元数	907	909	912	916	918	…
汉字数	34500	35000	35500	36000	36500	…
基元数	988	992	992	994	996	…
汉字数	51500	52000	52500	53000	53500	…
基元数	1046	1047	1050	1052	1055	…
汉字数	68500	69000	69500	70000	70500	
基元数	1085	1085	1085	1086	1086	

按抽样不同分为两子样本集合,第一个样本集(从1000个字符开始,取样间隔为1000个字符)用来建模,第二个样本集(从500个字符开始,取样间隔为1000个字符)用来拟合检验。

2.2　回归分析建立基元统计模型

汉字个数和汉字基元间是一种非确定的统计依赖关系,而回归分析就是研究这种关系的一种有力方法。因此文中采用回归分析来对样本集合进行建模,得到曲线拟合预测方程。

分析样本集一,得出基元个数随着汉字的增加成非线性增长,速率随着汉字的增长越来越慢。在可线性化一元非线性回归常用的回归曲线方程中,幂函数曲线最符合离散点图规律,幂函数方程和相应的化为线性回归的换元公式如下:

幂函数 $y = ax^b$,

令 $y' = \ln y, x' = \ln x, a' = \ln a$,

则有 $y' = a' + bx'$。

借助MINITAB统计软件,按照换元公式,将汉字个数 x 和汉字基元数目 y 分别输入到 $C1$ 和 $C2$ 列,再计算出 $\ln(C1)$ 和 $\ln(C2)$,分别放到 $C3(x')$ 和 $C4(y')$ 列,对 $C3(x')$ 和 $C4(y')$ 进行回归分析,得到如表4的结果:

表4　回归分析结果表

回归系数分析	自变量	系数	标准误差	T	P	
	常量	5.52684	0.01124	491.83	0.000	
	C3	0.131316	0.001096	119.82	0.000	
	S = 0.00834985 R-Sq = 99.5% R-Sq(调整) = 99.5%					
	回归方程: $y' = 5.53 + 0.131x'$					
方差分析	来源	自由度	SS	MS	F	P
	回归	1	1.0009	1.0009	14356.55	0.000
	残差误差	69	0.0048	0.0001		
	合计	70	1.0058			

分析表4,尾概率(拒绝假设检验的概率) P 为0.000,假设检验水平为 $\alpha = 0.01$, $p = 0.000 < \alpha = 0.01$,认为在 $\alpha = 0.01$ 这个检验水平下一元线性回归方程 $y' = 5.53 + 0.131x'$ 是有统计意义的,也即说明可线性化的一元非线性回归方程能够很好地反映汉字基元随着汉字数目增加的趋势,根据换元公式计算出幂函数系数 $a = e^{5.53} = 252\frac{1}{2}$, $b = 0.131$,那么一元非线性回归方程为:

$$y = 252x^{0.131} \tag{1}$$

2.3　回归模型的曲线拟合精度验证

为了验证得到的回归方程能够很好地拟合出基元库中基元数量随着汉字数量增加的变化规律曲线,采用MATLAB中的拟合工具箱Cftool中的Custom Equations自定义非线性拟合方程(1)来对样本集二的数据进行拟合,其中 x_2 是汉字个数, y_2 是基元数目,MATLAB参数曲线拟合基元库中基元数量随着汉字数量变化的散点图和拟合曲线如图2所示。

利用工具箱,分析得到的拟合吻合度指标中拟合系数为99.96%,调整后拟合系数不变,均方根误差为2.405。通过拟合吻合度分析,使用该曲线方程进行拟合,精度能达到99.96%,验证了回归分析得到的拟合方程(1)的可行性。

2.4　基元数目预测结果

依照(1)回归曲线模型方程,推断出 x 即汉字数量在100000个的时候,基元个数 y 的点预测是1139个,汉基元个数在95%的区间预测中为(1120,1159),99%的区间预测中是(1114,1166)。总计10万汉字,汉字基元的总数低于1159个的可能性为95%,低于

1166个的可能性达到了99%，也就是说，现在的汉字基元个数是1086个，预计最多只需要再增加80个左右的汉字基元就可以造出所有的10万汉字。相对于原来数量增加得很少，基元库可以基本维持相对稳定。

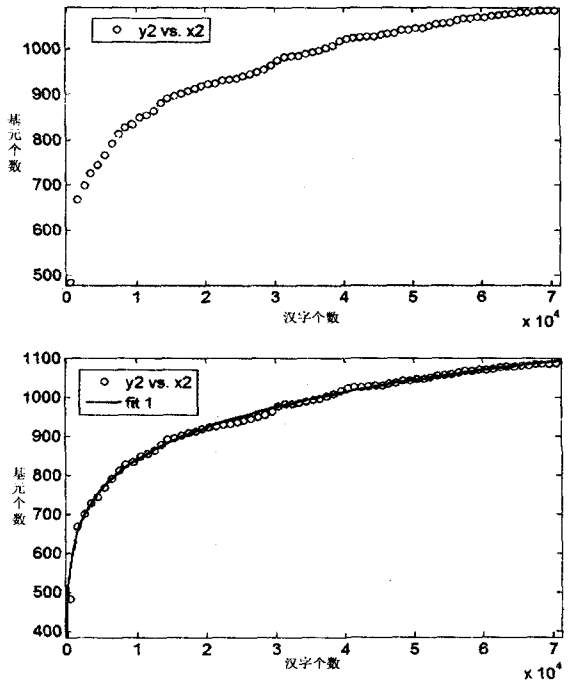

图2 GB18030-2005基元数随着汉字数变化的散点图(上)和拟合曲线(下)

3 结束语

文中对无字库造字系统中遴选出的汉字基元进行重复汉字实验，对汉字基元进行统计和分析，借助MINITAB软件对统计数据进行回归分析，分析其回归系数和方差，得到较为精确的回归方程 $y = 252x^{0.131}$。借助MATLAB仿真软件中cftool工具箱，用上述方程对得到的统计数据进行拟合，结果表明计算出的回归方程用来拟合基元离散数据精度很高，获得汉字数量与汉字基元的关系曲线比较准确。然后利用该回归方程预测和估计在10万汉字时汉字基元的数量。结果表明在99%的置信度区间内，全部十万汉字时基元的数量最多为1166个左右，在现有的基元数量上增长很少，也说明可以通过使用基元库取代字库来建立长期稳定的信息化标准。

参考文献：

[1] 卫红春.汉字信息处理技术发展概论[J].微机发展(现更名:计算机技术与发展)，1995,15(5):3-10

[2] 梁添才.基于认知机理的汉字智能造字研究[D].广州:华南理工大学,2008.

[3] Lai P, Pong M. Approaches to handle user-defined Chinese characters[C]//Proceedings of the International Conference on Computer Computing. [s. l.]:[s. n.],1994:235-241.

[4] 周浩华,蔡颖嘉.智能汉字库的研究[J].华南理工大学学报(自然科学版),1992,20(2):1-7.

[5] Lai Pak-Keung, Yeung Dit-Yan, Pong Man-Chi. A Heuristic Search Approach to Chinese Glyph Generation Using Hierarchical Character Composition[J]. Computer Processing of Oriental Languages,1996,10(3):307-323.

[6] 冯万仁,金连文.基于部件复用的分级汉字字库的构想与实现[J].计算机应用,2006,26(3):714-717.

[7] Pi Youguo, Liu Mingyou, Liao Wenzhi. Research on the Systemic Structure of Chinese Character Intelligent Formation [C]//APCIP2009. [s. l.]:[s. n.],2009:72-75.

[8] Liu Mingyou, Huang Jian, Pi Youguo. Acquiring the Mapping Knowledge of Basic Element in the Chinese Character Intelligent Formation[C]//ICICA2008. [s. l.]:[s. n.],2008:321-324.

[9] Pi Youguo, Liu Mingyou, Huang Jian. Knowledge representation based on semantic network in the Chinese character intelligent formation system[C]//ISKE08. [s. l.]:[s. n.],2008:785-790.

[10] Liu Mingyou, Huang Jian, Duan Chengsen, et al. Acquiring the Mapping Knowledge of Basic Elements Based on PSO in the Chinese Character Intelligent Information[C]//CISP09. [s. l.]:[s. n.], 2009:1-5.

[11] 丘志文.基于认知机理的汉字智能造字之汉字基元研究[D].广州:华南理工大学,2008.

[12] 刘明友.认知模式识别理论及无字库智能造字研究[D].广州:华南理工大学,2010.

(上接第32页)

[7] 侯志荣,吕振肃.基于MATLAB的粒子群优化算法及其应用[J].计算机仿真,2003,20(10):86-89.

[8] 王宏力,侯青剑.一种改进的粒子群优化算法及其仿真[J].自动化仪表，2009，30(7)：28-30.

[9] Shi Yuhui, Eberhart R C. A modified particle optimizer[C]// Proc. of the IEEE Conf. on Evolutionary Computation. [s. l.]:[s. n.],1998:69-74.

[10] 余 健,郭 平.基于改进的Elman神经网络的居家预测模型[J].计算机技术与发展,2008,18(3):40-45.

[11] 刘红梅,王少萍,欧阳平超.基于小波包和Elman神经网络的液压泵故障诊断[J].北京航空航天大学学报,2007,33(1):67-71.

[12] 王 涛,王晓霞.基于改进PSO-BP算法的变压器故障诊断[J].中国电力,2009,42(5):13-16.