GB18030 汉字的分形相关性研究*)

毛明毅[1] 陈志成[2] 李文正[1] 何华灿[3]

(北京工商大学计算机学院 北京 100037)[1] (清华大学信息技术研究院 北京 100084)[2]

(西北工业大学计算机学院 西安 710072)[3]

摘 要 GB18030是国家标准局新近颁布的最重要的汉字编码标准。本文从分形信息学的角度对GB18030汉字库中的27538个汉字的分形特性进行了研究。基于格分维理论,给出了汉字格测度的选取原则和格分维的计算方法。计算与统计表明:96.6846%的汉字的分形维数在[1.00, 1.50]之间,98.9469%汉字的分形相关性系数 R^2 值在[0.95, 1.00]之间,这表明汉字具有显著的分形特性。

关键词 GB18030,汉字,格分维,相关性

Study on the Fractal Correlation of GB18030 Words

MAO Ming-Yi[1] CHEN Zhi-Cheng[2] LI Wen-Zheng[1] HE Hua-Can[3]

(Computer Institute, Beijing Technology and Business University, Beijing 100037)[1]

(Research Institute of Information Technology, Tsinghua University, Beijing 100084)[2]

(College of Computer, Northwestern Polytechnical University, Xi' an 710072)[3]

Abstract GB18030 is the latest standard for encoding of Chinese words, which is constituted by the department of national standard of China. From the aspect of fractal informatics, this paper firstly studied the fractal characteristic of the 27538 Chinese words of GB18030 font libraries systemically. Based on the grid fractal theory, we give the principles of selecting grid measure for Chinese words and the procedure of calculating grid fractal dimension. The result is that and the grid fractal dimension of 96.6846% words are in range[1.00, 1.50], and the fractal correlation coefficient R2 of 96.6846% words are in range[0.95, 1.00]. This shows that the Chinese words have notable fractal characteristic.

Keywords GB18030,Chinese words,Grid fractal dimension,Correlation

1 引言

自20世纪末期以来,分形理论已经在几何拓扑[1,2]、生命系统[3]、图像处理[4]等众多领域得到了广泛应用。其中,盒子维、Hausdorff维[5]等在其发展与应用过程中起了重要作用。盒子维计算简便,但其测度的选取具有较强的经验性;Hausdorff维在数学理论上是完备的,但其求解繁琐,缺乏可计算性。文[6]在引入"参考格空间"和单位格测度"的基础上,提出了一种新的维数理论——格分维理论,它给出了一种适用于任意图形图像的分形维数计算方法,目前已经得到良好的应用[7,8]。

国家标准GB18030-2000《信息交换用汉字编码字符集基本集的扩充》[9,10]是我国继GB2312-1980和GB13000-1993之后最重要的汉字编码标准,是未来我国计算机系统必须遵循的基础性标准之一。国家标准局制定了针对产品的《GB18030标准符合性检测规范》,同时开发了GB18030基本点阵字型库[11],简称GB18030汉字库。

利用格分维理论来探讨汉字的分形特性,测度的选取不再凭经验,而是依据一定的原则和步骤进行,这可以消除分形维数的主观不确定性,结果相对客观可靠。基于此,本文研究GB18030大字库汉字的分形特性,编写程序计算了27538个汉字的格分维,进行了相关性分析。

2 汉字图像与格分维理论

2.1 汉字字模与字图像

GB18030汉字编码标准的编码空间约为160万码位,目前已编码的字符约2.6万。国家标准的字符集是以汉字库的形式提供的。汉字库通常是二进制格式文件,每个汉字在库中有确定的位置和大小。文[20]给出了GB18030大字库的解读程序。以16点阵为例,每个汉字用16×16的点阵来存储,每个点用一个二进位1或0来表示。如图1表示"大"字的字模。

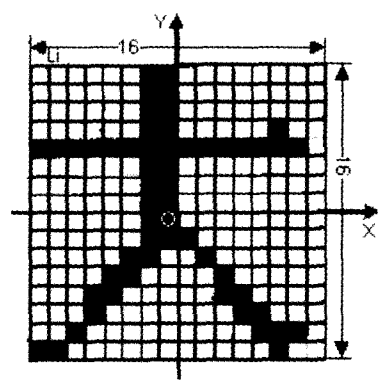


图1 16点阵的汉字字模

*)国家自然科学基金项目(编号:60273087)。**毛明毅** 博士,主要研究方向为人工智能与分形技术;**陈志成** 博士后,主要研究领域为分形理论与嵌入式系统;**李文正** 教授,主要研究方向为网络智能;**何华灿** 教授,博导,主要研究方向为人工智能基础及其应用、泛逻辑学与不确定性推理。

定义 1(字图像) 汉字字模可以看成是一幅黑白图像，存 1 的点表示有图案，存 0 的点仅代表背景色。为便于后文叙述，本文称此类图像为“字图像”。因此，每个汉字都唯一对应了一幅字图像，反之亦然。

针对字图像，可以利用分形理论来计算其分形维数，考察汉字的相关性质。

2.2 格分维理论的基本思想

关于计算图形图像的分形维数计算方法，传统的有盒子维和 Hausdorff 维等，但是它们存在如下缺陷：

(1)盒子维计算简便，但其测度的选取具有较强的经验性和主观性。因此，对于同一个图像，不同的研究人员计算所得的盒子维通常都不相同。由于缺乏统一的衡量标准，计算结果不具有可比性。

(2)Hausdorff 维在数学理论上被证明是完备的，但是由于涉及到最大最小极限值的求解，因此在实际上很少得到真正的应用，通常都是求取其近似值。因此，Hausdorff 维的最大缺陷在于：求解繁琐，缺乏可操作性。

(3)通过分析盒子维、Hausdorff 维以及现有的其它一些分形维数的计算方法，发现分形维数不统一的根本原因在于：计算过程中缺乏可参照的分形空间和单位测度。

本文利用新近提出的格分维理论来计算汉字的分形维数，详见文[6]。格分维理论从定义上和分形维数的计算方法上克服了上述缺陷，其基本思想在于：引入了“参考空间”和“单位格测度”的概念，从根本上彻底消除了由于测度选取的不确定性所带来的分形维数的不确定性，给出了统一的任意图形图像的分形维数计算方法。对于同一个图像而言，任何人的计算结果都是相同的，具有可计算性和可比性。

3 汉字图像的格分维计算方法

3.1 汉字图像的测度选取原则

利用格分维理论来求汉字的分形维数，把汉字看成一幅几何图像，给出测度选取的三个原则：

(1)确定参考格空间。必须把待研究的分形对象纳入某一确定维数的参考格空间中进行考察。设参考格空间的维数为 D_0，则其中的分形对象的维数应该在$[0,D_0]$之中，D_0 为实数。

(2)设对 $H\times W$ 的图像求分维。根据 H、W 大小选定最小测度与最大测度。其间有 n 个测度，利用程序从 n 个测度中选择 $m(m>2)$ 个测度组成测度序列 L_i，这会有 C_n^m 个序列。在假设整个图像全部有图案的条件下，求出对应于各个序列的分形维数。

(3)选取参考格空间的测度序列，应满足：①由所选测度序列求得的分维等于或十分接近于 D_0；②在 $m>2$ 且满足①的条件下，取包含元素个数较多且各个元素较为分散的测度序列作为参考格空间的测度序列。

定义 2(本征测度序列) 由上述三个原则选出的测度序列，称为维数为 D_0 的分形空间的本征测度序列。如计算嵌入其中的分形对象的分形维数，需用此本征测度序列去进行测量。

3.2 字图像的格分维计算方法

根据格分维理论，维数的计算公式为 $D_G(F)=f(S,\delta,F)$[6]。结合上述字图像测度的选取原则，我们给出计算汉字的格分维的方法如下：

(1)选取单位格大小 δ 为一个像素，这是计算机图像的最小分辨单位；

(2)选取参考格空间 S 为其本征格空间，设其维数 D_0。字图像即为对象集 F。对于本文的字图像，笔者纳入 D_0 为 2 的参考格空间中进行研究，其中汉字的分维可以分布在[0,2]之间，这打破了传统的认为二维图像的分维应该在[1,2]之间的观点；

(3)根据字体与字号(高度(H)与宽度(W))，利用程序求得参考格测度集 L，进而得到本征测度序列 L_i，$i=1,2,\cdots m$；

(4)利用 L_i 值，分别构造面积为 L_i^2 的方格，得到子空间序列；

(5)用子空间序列去覆盖整个字图像，则可以得到对应于不同 L_i 的格子数(N_i)；

(6)对 L_i 值进行单位化，并计算 $\ln(1/L_i)$值和 $\ln(N_i)$值。

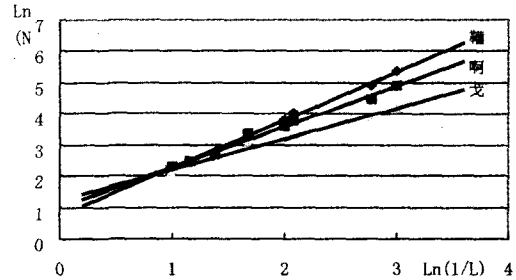
(7)以 $\ln(1/L_i)$为横坐标、$\ln(N_i)$为纵坐标在二维坐标系中描点作图，并进行线性回归分析，所得直线的斜率即为字图像的分形维数 D，此即为“格分维”。

4 汉字的分形特性研究

研究汉字分形特性的重要内容之一是确认汉字是否具有分形特性。根据分形事物的标度不变性，所拟和 $\ln(N)\sim\ln(1/L)$直线的线性相关性程度评估值 R^2 表示了汉字是否具有分形特性或具有分形特性的程度大小，R^2 越接近于 1，则汉字越具有分形特性。一般而言，在统计数学中[12]，R^2 值高于 0.95 即可认为具有显著的线性相关性，R^2 值在 0.90～0.95 之间亦可认为具有一定的线性相关性。

4.1 汉字格分维的计算与分析范例

这里依照汉字笔划的多少与繁简程度选取“鞾”、“啊”、“戈”三个具有代表性的汉字来分析其分形特性。其中，“鞾”是繁体字，其笔划较多；“啊”是字库中的第一个汉字，其笔划比鞾要少，“戈”字最简单。图 2 是相应的 $\ln(N)\sim\ln(1/L)$直线，其中列出了鞾、啊、戈三字的线性回归直线方程。由图可知：鞾、啊、戈三字的 R^2 值分别为 0.9964、0.9899、0.9834，故其具有显著的分形特性。



字	回归直线方程	R^2值	分维
鞾	$y=1.5543x+0.6940$	0.9964	1.5543
啊	$y=1.3044x+0.9891$	0.9899	1.3044
戈	$y=0.9922x+1.2077$	0.9834	0.9922

图 2 鞾、啊、戈三字的 ln(N)～ln(1/L)直线图

4.2 汉字分形特性相关性分析

本文针对 GB18030 宋体、16 点阵字库中的汉字进行计算，计算量为 27538 个汉字。程序计算时既输出了汉字的分形维数，同时输出了其线性相关性系数 R^2 值。其统计结果见图 3、图 4 及表 1 所示。

(下转第 207 页)

化参数优化所得的超分辨率图像。从结果看双线性插值生成的高分率图像图 1(b)明显模糊。图 1(c)的结果明显清晰，也明显比图 1(a)清晰。说明用 Arnoldi 过程方法确定的正化参数优化所重建的超分辨率图像结果明显好。

结论 本文使用 Tikhonov 正则化方法和梯度最速下降优化方法求解超分辨率图像，我们提出基于 Arnoldi process 的正则化参数快速确定方法。它只需要一次使用 Arnoldi process 算法，而计算每个正则化参数只需少量的计算即可。理论和实验证明该方法计算正则化参数是快速和有效的，对大尺度线性超分辨率图像复原问题的参数估计是有效的工具。

参 考 文 献

1 Tsai R Y, Huang A K. Multi-frame image restoration and registration. In: As advances in Computer Vision and Image Processing, 1984, 1:317～339

2 Hansen P C, O'Leary D. The use of the L-curve in the regularization of discrete ill-posed problems. SIAM J Sci Comput, 1993, 14(6): 1487～1503

3 Molina R, Vega M, Abad J, et al. Parameter Estimation in Bayesian High-resolution Image Reconstruction with Multisensors. IEEE Trans on Image Processing, 2003, 12(12):1655～1667

4 Park S C, Park M K, Kang M G. Super-resolution Image Reconstruction: A Technical Overview. IEEE Signal Processing Magazine, 2003. 21～36

5 Golub G H, Van Loan C F. Matrix Computations. The Johns Hopkins University Pree, 1996

6 Elden L. Algorithms for the Regularization of Ill-conditioned Least Squares Problems. BIT, 1997,17:134～145

7 张新明. 超分辨率图像复原的研究:[博士论文]. 北京:北京工业大学, 2002

(上接第 184 页)

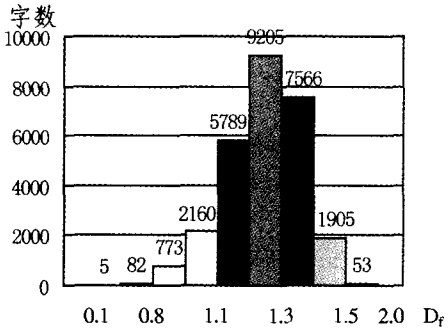


图 3 GB18030 汉字的格分维计算

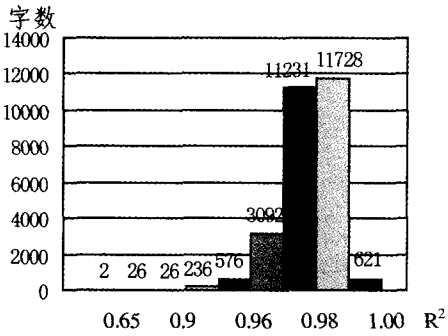


图 4 GB18030 汉字的相关性分析

表 1 GB18030 汉字的格分维 D_f 与相关性系数度 R^2 统计表

D_f 值区间	字数	百分比%	R^2 值区间	字数	百分比%
[0.10,0.50)	5	0.0182	[0.00,0.65)	2	0.0036
[0.50,0.80)	82	0.2978	[0.65,0.85)	26	0.0944
[0.80,1.00)	773	2.8070	[0.85,0.90)	26	0.0944
[1.00,1.10)	2160	7.8437	[0.90,0.95)	236	0.8569
[1.10,1.20)	5789	21.0219	[0.95,0.96)	576	2.0915
[1.20,1.30)	9205	33.4265	[0.96,0.97)	3092	11.2281
[1.30,1.40)	7566	27.4748	[0.97,0.98)	11231	40.7836
[1.40,1.50)	1905	6.9177	[0.98,0.99)	11728	42.5884
[1.50,2.00]	53	0.1925	[0.99,1.00]	621	2.2551
合 计	27538	100	合 计	27538	100
[1.00,1.50)	26625	96.6846	[0.95,1.00]	27248	98.9469

由图 3、图 4 及表 1 可知：大多数汉字的格分维集中在[1.00, 1.50]之间，分形相关性系数 R^2 大于 0.95。其中，作者对 R^2 值小于 0.90 的汉字作了详细分析，其中 8 个汉字是空格或空白字符图形，有 5 个是一级字库与二级字库之间的间隔字符，其余部分的字图像仅仅在点阵字模的"边界"上有少量为"1"的像素点，其线性相关性较弱。但总体而言，96.6846%汉字的分形维数在[1.00, 1.50]之间，98.9469%汉字的 R^2 值在[0.95, 1.00]中。因此可以说，GB18030 汉字具有显著的分形特性，在很大程度上具有分形图像的标度不变性，可以利用分形理论来研究汉字的相关性质。

结论 本文工作的特色在于：(1)研究对象是国家标准局新近颁布的 GB18030 大汉字库中的 27538 个汉字的分形特性；(2)所用理论是作者在文[6]中新近提出的格分维理论及其计算方法；(3)所得结论是 96.6846%汉字的分形维数在[1.00, 1.50]之间，98.9469%汉字的分形相关性系数 R^2 值在[0.95, 1.00]之间，这表明汉字具有显著的分形特性，可以利用分形理论研究汉字的相关性质。

汉字本身是一个复杂的图形，而且是在逐渐演化的图形，对汉字的研究是一个复杂的课题。本文工作仅是从分形维数的角度来研究，文中涉及的思想方法和计算数据对于相关领域的研究者具有一定的借鉴作用和参考价值。

参 考 文 献

1 Falconer K J. The Geometry of Fractal Sets [M]. London: Cambridge University Press, 1985

2 Mandelbrot B. Fractal Geometry of Nature [M]. Freeman, San Francico, 1982

3 叶竹秋，林跃鑫. 生命科学中的分形研究[J]. 自然杂志，2001, 23(2): 87～90

4 陈衍仪. 图像压缩的分形理论和方法[M]. 北京：国防工业出版社，1997

5 Grassberger P. Generalizations of the Hausdorff Dimension of Fractal Measures [J]. Phys Lett, 1985, 107A: 101～105

6 陈志成. 复杂系统中分形混沌与逻辑的相关性推理研究[D]:[学位论文]. 西安：西北工业大学，2004

7 陈志成，何华灿，毛明毅，等. 基于格图象的康托集分维与泛逻辑运算[J]. 计算机科学，2004, 31(4): 92～95

8 毛明毅，何华灿，陈志成，等. 分形图像的泛逻辑运算模型[J]. 计算机工程与应用，2004, 40(2):23～56

9 国家质量技术监督局. GB18030－2000，信息交换用汉字编码字符集基本集的扩充[S]，2001

10 林宁. 关于 GB18030 汉字编码标准集[N]. 中国计算机报，2001, 54

11 陈志成，何华灿，毛明毅. GB18030 字库的解读与压缩封装程序设计[J]. 计算机工程与应用，2002, 38(18):119～129

12 沈恒范. 概率论与数理统计教程[M]. 北京：高等教育出版社，1995