维普资讯 http://www.cqvip.com

关于甲骨文象形码输入法的编码原理研究*)

Study for Design Principle of the Input Method of JiaGuWen Symbol Code

胡金柱　肖　明

(华中师范大学计算机科学系　武汉430079)

Abstract This paper makes use of the theory of Fuzzy information analyser, researchs and sets up the fuzzy information model for the Input method of Jia Gu Wen symbol code. It's design principle of input method of Jia Gu Wen Symbol code. On this basis, we put to use of the entropy theory in informatics, research and analyse the efficiency and scientificalness of the coding of Jia Gu Wen symbol code. And the research of other ancient's word information processing or other Asian's word information processing can benefit from the design principle of the Input method of Jia Gu Wen symbol code.

Keywords Chinese information processing, Jiaguwen, Symbol code, Input method

1. 前言

甲骨文是我国最早的文字。迄今为止，广大学者已从全部甲骨文刻辞中整理出五千多个具有代表性的甲骨文字[1]。经研究考释，发现甲骨文字形独特，是一种典型的象形文字，具有"图画"的某些特点。但因年代久远，所以对于甲骨文无论从音或是义上人们都很难完全把握，多年来，一直没能对甲骨文进行有效地编码。事实上，任何一种文字都自成体系，有一套确定的构成规律，它们的构形部件尤其如此。甲骨文也不例外，它的构形部件亦有一定的规律。针对该特点，在研究中吸取现代编码思想，运用模糊数学和面向对象 Petri 网方法研究甲骨文的部件(字根)形成和码元的确定规则，分析甲骨文编码的效率和科学性，为现有的5000多个甲骨文字进行科学编码提供理论依据。

2. 甲骨文信息处理中的模糊信息模型

中国社科院考古所编写的《甲骨文编》中记录了大约4万个不同形体的甲骨文，它们代表五千多种不同含义的甲骨文字，其中与后世文字有联系并能够辨识确定的仅有2千多个字，其余的是一些表示人名、地名的文字。分析发现，在甲骨文字之间存在大量"亦此亦彼"，处于差异中介过渡状态的模糊现象。因此，本文运用模糊信息分析理论，建立起甲骨文信息处理的模糊信息模型[2]。

2.1 甲骨文编码的模糊信息模型

甲骨文是具有完整体系的早期汉字，其构形部件——甲骨文字根也带有一定的象形性、确定性和稳定性。我们在研究甲骨文及字根时，考察发现甲骨文字根大致可以划分为500多个，形态大致有32种。但字根之间在形态上存在"像"与"不像"的模糊性，因此完全可以采用模糊聚类技术[3]将500多个甲骨文字根按其形态定义为如下32个模糊集合：

x_1：表示字根对模糊集"$\underline{A}_1$"的隶属度；x_2：表示字根对模糊集"$\underline{A}_2$"的隶属度；……；x_{32}：表示字根对模糊集"$\underline{A}_{23}$"的隶属度。

其中，$\underline{A}_1$＝{金字塔形或与字母"A"相像的字根}，$\underline{A}_2$＝{目型或与字母"B"相像的字根}，$\underline{A}_3$＝{月牙形或与字母"C"相像的字根}，……，$\underline{A}_{32}$＝{半挑形或与字母"P"相像的字根}等共32种。

由于甲骨文字根不像现代汉字那样具有很强规则的笔划(如：丶、一、丨、丿、乚)，也不像甲骨文字那样具有较完整的拓扑结构(端点、叉点、孔、块)，甲骨文字根大部分已被抽象成只有简单的"绘画"般的线条，故不能像汉字那样从规则笔画中来提取特征，或像甲骨文字那样从拓扑结构中来提取特征。因此，我们考虑用各字根对这32个模糊集的隶属度来作为该字根的特征。即：令 $u_i=(x_{i1}, x_{i2}, \cdots, x_{i32})$ 表示第 i 个待分类的甲骨文字根，其中 x_{ij} 为 u_i 对第 j 个模糊集的隶属度。

由于甲骨文字根是定性的、离散性的，无法直接对它们进行量化，于是采用专家评估法可以建立如下相似模型：

$$S_{ij}=\frac{\text{认为第 } i \text{ 个字根应属于第 } j \text{ 个模糊集的人数}}{n}$$

其中 n 为参加评估的专家人数。

利用上述模型可对500多个甲骨文字根进行完整地定位。例如随机抽取10个字根，求得其 S_{ij}($i=1,2,\cdots 10$, $j=1,2,\cdots,32$)的值如表1所示。

2.2 模糊聚类分析

(1)建立甲骨文字根对象的 F 相似关系：设 $U=\{u_1,u_2,\cdots,u_n\}$ 为待分类的全体字根，其中每一待分类的字根对象为一组数据：$u_i=(x_{i1},x_{i2},\cdots,x_{im})$。于是采用数量积法可建立 u_i 与 u_j 之间的相似关系 $R(u_i,u_j)=r_{ij}$ 如下：

$$r_{ij}=\begin{cases}1 & \text{当 } i=j\\ \dfrac{1}{M}\displaystyle\sum_{k=1}^{m}x_{ik}\cdot x_{jk} & \text{当 } i\neq j\end{cases}$$

其中 M 为 $M\geqslant\max_{i,j}\left(\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\right)$ 的正数。

(2)建立字根对象的相似矩阵：根据数量积法公式，计算求得 $\max_{i,j}\left(\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\right)=0.64$，(其中 $m=32$)，故可取 $M=0.7$，计算可得相似矩阵 R 如表2所示。

(3)画出聚类图：直接将 R 图中权重(r_{ij})不低于 λ 的路联起来，在一条路上的元素就是一类。取 λ 从1到0，可得所有的分类。例如：取 $\lambda=0.8$，权重不低于0.8的路如图1所示。共分为六类：{1,4,7,9}，{6,8}，{2}，{3}，{5}，{10}。

*)本文为国家重点实验室和湖北省自然科学基金资助研究项目。

维普资讯 http://www.cqvip.com

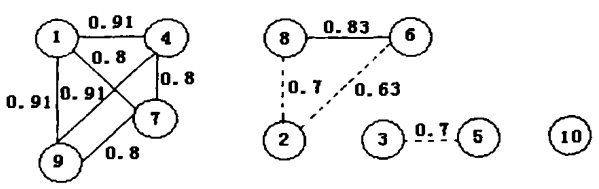

图1

表1

样本 i / 变量 j	1	2	3	4	5	6	7	8	9	10
x_1	0.8		0.2	0.8			0.7		0.8	
x_2		0.1					0.3			
x_3		0.6				0.7		0.8		
x_4						0.1		0.2		
$x_5 \sim x_8$										
x_9									0.1	
$x_{10} \sim x_{12}$										
x_{13}		0.1				0.1				0.2
x_{14}				0.1						
x_{15}	0.2		0.1							
x_{16}									0.1	
$x_{17} \sim x_{20}$										
x_{21}					0.2					
x_{22}			0.7	0.1	0.7					
x_{23}		0.1				0.1				0.8
x_{24}										
x_{25}					0.1					
$x_{26} \sim x_{30}$										
x_{31}		0.1								
x_{32}										

表2 相似矩阵

r_{ij}	1	2	3	4	5	6	7	8	9	10
1	1	0	0.26	0.91	0	0	0.8	0	0.91	0
2		1	0	0	0	0.63	0.04	0.7	0	0.2
3			1	0.33	0.7	0	0.2	0	0.22	0
4				1	0.1	0	0.8	0	0.91	0
5					1	0	0	0	0	0
6						1	0	0.83	0	0.2
7							1	0	0.8	0
8								1	0	0
9									1	0
10										1

若取 λ=0.6，则权重不低于0.6的路在上述的路上把{2}和{8}联起来，{3}和{5}联起来。这时，10个字根可分为四类（即四个码元）。{1,4,7,9}，{2,6,8}，{3,5}，{10}，该聚类图如图2所示。

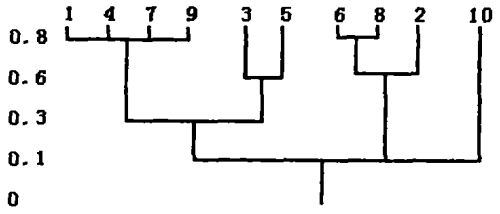

图2

由于甲骨文字根有500多个，采用上述聚类方法对全部字根一次聚类将十分困难。因此，将500多字根随机分成10组，在各组中分别进行聚类处理，从而将这500多个甲骨文字根分成32类，称之为码元。再用键盘上的32个字符（包括25个字母和7个数字：1、2、3、4、5、8和9）对应32个码元，其对应关系如图3所示。

将 $B_i(i=1,2,\cdots,m)$ 看作是一个特殊的元素，可以用它来定义它对 $\underset{\sim}{A}_j=1,2,\cdots,32)$ 的隶属度 $\sigma(B_i,\underset{\sim}{A}_j)$：

设 λ 值取定后，B_i 中的元素为 $\{b_1^{(i)},b_2^{(i)},\cdots,b_k^{(i)}\}$，映射 σ：□x□→[0,1]，其中□=$\{B_1,B_2,\cdots,B_m\}$，□=$\{\underset{\sim}{A}_1,\underset{\sim}{A}_2,\cdots,\underset{\sim}{A}_{32}\}$，于是定义

$$\sigma(B_i,\underset{\sim}{A}_j)=\frac{1}{k_i}\sum_{t=1}^{k_i}u_{\underset{\sim}{A}_j}(b_t^{(i)})$$

若取 $\sigma(B_i,\underset{\sim}{A}_t)=\max_j\sigma(B_i,\underset{\sim}{A}_j)$，则可以将 B_i 归入 $\underset{\sim}{A}_t$ 中。由此，可将500多个甲骨文字根与32个码元建立起完整的对应关系。

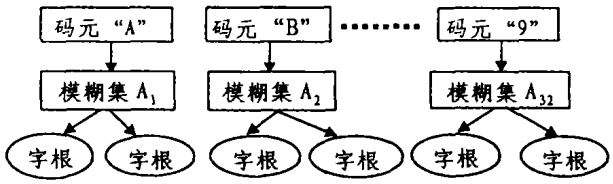

图3

3. 甲骨文字编码的熵和效率

3.1 甲骨文字编码的熵

利用热力学中表示热系统无序程度或不确定性的"熵"(Entropy)，来描述系统信息缺乏程度的度量，则称之为"信息熵"。对于具有 n 个信息元素的信息集合，熵的一般表达式为：

$$H=-\sum_{i=1}^{n}P_i\log P_i \qquad (1)$$

式中，P_i 表示 n 个状态中的第 i 个状态的发生概率，H 代表消除系统不确定性所需的信息量，即以"位"为单位的负熵，H 越大则系统信息越不确定。

熵具有一条基本性质：等概率分布具有最大熵，即在(1)式中，当等概率情形 $P_i=1/n(i=1,2,\cdots,n)$ 时，有：

$$H_{max}=-\sum_{i=1}^{n}(1/n)\log(1/n)=\sum_{i=1}^{n}(1/n)\log(n)=\log n \qquad (2)$$

其他情况下，其熵均为 $H<\log n$。

信息编码可分为等长码与不等长码两种编码方式。由熵的性质可知，等长码符合最大熵的条件，但编码效率最低；为了压缩信息存储空间和节省信息传输时间需要提高信息的编码效率，采用不等长码编码方式——Huffman 编码（最佳编码）。

对于5000多个甲骨文字，由式(2)求得采用2进制码时，其平均熵为：

$$H(\text{甲骨文字等长码})=\log_2 5000\approx 13\text{位} \qquad (3)$$

取不等长码则有：

$$H(\text{甲骨文字不等长码})=-\sum_{i=1}^{n}P_i\log_2 P_i$$

式(1)中 n 为参加编码的甲骨文字的总数，P_i 为第 i 个甲骨文字的使用频度。由于甲骨文字的使用频度分布很不均匀，

维普资讯 http://www.cqvip.com

所以比等长码13位要短些。熵值小，这意味着可用较短的码位来进行编码。

3.2 甲骨文字的编码效率

在甲骨文编码时，我们采用了直观的字母数字混合符，所以对甲骨文字编码的码元数大于2。现设甲骨文字符集有 n 个甲骨文字，使用 m 个码元对它进行编码，第 i 个甲骨文字的码长为 l_i，用 P_i 表示其根据汉字的使用频度推测的甲骨文字使用频度，根据 Huffman 最佳编码方法有[5]：

$$\log_m(1/P_i)<l_i<\log_m(1/P_i)+1 \quad (4)$$

其中 $\log_m(1/P_i)$ 是第 i 个甲骨文字的信息量。

由于一个甲骨文字的使用频度不一定刚好等于 m 的负整数幂，所以取上述不等式形式。下面用 P_i 乘以式(4)，并遍历对其求和得

$$\sum_{i=1}^{n}P_i\log_m(1/P_i)<\sum_{i=1}^{n}P_il_i<\sum_{i=1}^{n}[P_i\log_m(1/P_i)]+\sum_{i=1}^{n}P_i \quad (5)$$

根据 Huffman 编码公式 $r=L/H-1$，其最短平均码长为：$l_{min}=\sum_{i=1}^{n}P_il_i$。式中最后一项的概率和为1，而 $\sum_{i=1}^{n}P_i\log_m(1/P_i)$ 就是编码甲骨文字字符集以 m 为底的熵。如用 H_m 表示，则不等式(5)为：

$$H_m<l_{min}<H_m+1$$

其中的 H_m 就是 Huffman 最佳编码长度的理论下限，它指出任何文字编码方案的平均码长都不可能达到此下限。例如根据汉字的使用频度，在不考虑语言符号出现概率相互影响时，当汉语文本中的汉字容量 n≥12370时，得到 $H_2=-\sum_{i=1}^{n}P_i\log_2P_i=9.65$位[6]。

当编码码元数 m 增加时相应的最佳编码长度随之缩短。计算结果如表3所示。

表3

m	l_{min}	
	下限值	上限值
10	3	4
26	2.05	3.05
32	1.97	2.97
47	1.71	2.71

根据前面对甲骨文字根的模糊聚类分析，形成了32个码元，即 m=32。从表中可以看出，对甲骨文编码的平均码长大致接近于3，差不多处于最佳码长的上限左右。由于甲骨文字根有500多个，仅由32个码元表示，会出现同一码元的字根数较多，造成重码率较高的现象。按照最佳码长编码，试图获得一个编码效率高、重码率低的理想甲骨文编码是十分困难的。为了减小重码率，甲骨文编码采用增加编码规则，即在末码上按其结构类型取码，使得某些甲骨文编码的码长为4。即甲骨文编码的理论最佳码长为3，而实际编码时最多为4位码长。

结束语 本文吸取现代编码思想，运用模糊信息分析理论，建立了甲骨文信息处理的模糊信息模型。利用模型分析甲骨文的部件(字根)和码元的确定规则，使用32个字符(25个英文字母和7个阿拉伯数字)作为码元，与甲骨文中的500多个字根相对应，实现了一字一码的编码方案，首次对甲骨文进行了有效地编码。运用信息论中的熵理论，分析了这种编码的效率和科学性，得出甲骨文编码的最佳码长大致接近于3，从而为5000多个甲骨文字进行科学编码提供了理论基础。

在计算机甲骨文象形码输入法(含甲骨文字库)的基础上，进行象形逻辑模型和象形对象演算系统的研究；针对当前国内外关于汉字识别与密码学的研究现状，研究一种全新的古文字识别方法和现代网络环境下的加解密技术。这些都是需要进一步深入研究和探索的问题。

参 考 文 献

1 徐中舒. 甲骨文字典. 四川辞书出版社，1988
2 肖明，胡金柱，陈增照，舒忠梅，谭支鹏. 甲骨文象形码编码原理初探. 见：1998中文信息处理国际会议论文集，清华大学出版社，1998. 80～83
3 胡金柱. 模糊决策与决策支持. 华中师范大学出版社，1990
4 黄崇福，王家鼎. 模糊信息分析与应用. 北京师范大学出版社，1992
5 徐松，胡金柱. 甲象文象形码输入法的实现. 华中师大学学报(自然科学版)，1995，29(3)：299～302
6 陈一凡，胡宣华. 汉字键盘输入技术与理论基础. 清华大学出版社，广西科学技术出版社，1994
7 Corp A，Adobe Technical Note #5176：The Compact Font Formats Specification，1998，3
8 Janousek V，Tomas Vojnar State Spaces of Object-Oriented Petri Nets. In：Proc. of the Workshop on Concurrency within the MFCS/CSL'98 conf. organized，University of Brno，Czech，1998
9 Johnson R A，Hardgrave B C. Object-Oriented methods. current practices and attitudes，The Journal of Systems and Software，1999，48(1)：5～12

(上接第138页)

限，不过被选取的这几种情况具有典型性和一定的代表性，基本反映了具有统计意义的性能变化趋势和区别。更精细的调整有关参数或进一步改进 pool_manage()函数的算法，一定还会得到更好和更准确的分析结果；如果在多处理器系统上，多线程动态 POOL 技术将会显示出更优越的性能。

这里仅以 TCP Server 的设计为例，分析和研究了多线程动态 POOL 的设计技术，并且可以表明它对提高 Server 的性能有明显的作用。线程 POOL 的思想应该说来自于进程 POOL 的观念，所以，在进程 POOL 可以应用的很多地方都可以采用线程 POOL，并可减少开销和提高性能。不仅如此，在很多数据库 Server 和应用 Server 的设计中，尤其是具有多处理器或多机系统的环境，深入研究和应用动态线程 POOL 技术都将可能在不同程度上得到性能的提高。

参 考 文 献

1 MDMA - Multithreaded Daemon for Multimedia Access. http://www.ibiblio.org/mdma-release/mdma.html
2 An Approach for Constructing Web Enterprise Systems on Distributed Objects. http://www.redbooks.ibm.com/redpapers/pdfs/redp0025.pdf
3 Distributed Systems & IP Services. http://www.multithread.co.uk/
4 Multithread Programming Technology. http://www.npac.syr.edu/projects/pcrc/cpswt-e/toc/node1.htmlrg
5 Installed files for Apache Web Server，:www.linuxdoc.org
6 W Richard Stevens，UNIX networking Programming Volume I 2nd ed，networking APIs：Sockets and XTI，1998 by Printice Hall