PROCEEDINGS OF
1983 INTERNATIONAL CONFERENCE ON CHINESE INFORMATION PROCESSING

中文信息处理国际研讨会
论 文 集

中国中文信息研究会和联合国教科文组织联合举办

1983年10月12日至14日于北京

CO-SPONSORED BY CIPSC & UNESCO
BEIJING, PEOPLE'S REPUBLIC OF CHINA
OCTOBER 12-14, 1983

数据加载失败，请稍后重试！

序

一九八三年北京中文信息处理国际研讨会(ICCIP)是由中国中文信息研究会(CIPS of CHINA)和联合国教科文组织(UNESCO)共同举办的。

这次会议得到了许多国家和地区同行们的支持。论文集收编了70余篇文章，涉及到：汉字信息处理系统、汉字计算机输入输出技术、汉字文字识别与汉语语音识别、汉字输入编码以及有关基础理论等多方面的学术问题。我希望并相信，这个论文集将有益于中文信息处理技术的研究和推动在使用汉字地区、国家对中文计算机的广泛应用。

中国中文信息研究会理事长
中文信息处理国际研讨会主席

钱伟长

一九八三年九月十二日

编者注：论文集按文章收到的时间先后排序。

目　录

中文信息的处理

基本笔划方案的科学原理

中文电脑打字机的键盘设计

澳大利亚　卢遂现（Dr.S.Y.LO）

1、 科学原理

在科学上一条重要原则就是要分解事物到最基本的单元。例如原子就是物质的单元；0和1是计算机里的两个基本单元。有了正确的单元之后，一切复杂的事物便可以变简单了。在文字方面也一样，特别是汉字，它不能象英文那样由字母构成，必须另外找其最基本的组成单位。寻求汉字基本组成的单元便是寻求解决一切和汉字有关问题的钥匙。

有人认为应该用部首，但部首不完全，而且部首的数目也有二百多个。一方面太多，另一方面有了部首也不能全面地代表所有汉字。任何查过现有中文字典的，都会体验到这方面的不完全性和不唯一性。所谓唯一性就是说，一些虽然完全由部首组成的汉字也有不同的划分方法，没有统一的部首组成准则，便会在输入汉字时，必然带有重码问题或要另外创造一个任意的不客观的规则问题。

既然汉字不能由部首组成，能不能说汉字由点组成呢？汉字完全可以由点组成，现有的点阵打印机能打汉字便完全证明了可能性。但一个汉字最少要有16×16的点阵，用电脑作输出还可以，要人用点组成一个字或牢记汉字中点的位置和数目、输进这些数据是绝对不切实际的。

在点和部首当中能不能找到一些组成汉字的单元呢？答案是能够。这就是线了。平面形的汉字可以分解为由线点组成，而线又可分解为点。西方文字的字母也都可以由点、线组成，因此点、线是一切文字的基本组成部分。

点、线怎样分类呢？点在几何学上，只有一种，不能再分。线却有多种，主要分直线和曲线两种。英文字不能没有曲线，但汉字刚巧有这特点，完全可由直线组成，由此，省去了许多不必要的再分类。

直线可分长短和方向。在输入汉字代码中，什么是最少、最基本的，而又最容易记牢的呢？就是不计长短，而只有0°、45°、90°、135°的直线了。

所以从几何上，汉字最基本的组成部分便是：·一丨／＼五种。不能再简单了。当然可以把类别弄得更细更复杂一点，包括更多种类型的直线，甚至曲线。然而，由于这五种是最基本的，其它复杂的组成部分都可以归纳到里面去。

有了汉字最基本的几何单元，便需要把它们化作数字，以便计算机处理，将·一丨／＼称为12345，五个数字。汉字代码基本便完全可由这五个数字组成。因此可以说我们的方法是基本笔划法，以别于其他各汉字的代码方法。

基本笔划法的基本性也表现于代数上。我们知道，计算机的最基本语文是用0、1两个数字组成，不管千变万化的各种运算，机器语文最后都化作0、1两个数字的组合。有时高级计算机语文用alpha nu-meric的系统，26个拉丁字母加10个阿拉伯字母。有时在计算机的内储用16位的数字，用的单元是0,1,2,3,4,5,6,7,8,9,加上A,B,C,D,E,F共十六个符号。当然在普通数学运算上，还是用0,1,…,9的十进数。不管十进，十六进，还是用拉丁字母，最后都可以，而必须化成0,1,两个数字的组合。这样，一切计算机才有共同的工具。因此，最简单、最基本的0和1却是最全面、最有用的。有了最基本的0和1，毫不妨碍其他更复杂的符号和代号的发展，恰恰相反，有了最基本的0和1，才可能有花样极多的各种高级计算机语文。

同样我们的汉字代码用了1,2,3,4,5,五个最基本组合，才是最简单的共同代号。在这个简单代号的基础上，当然可以作出花样万千的各种新代号，适应不同情况下的不同需要。但是这些都不是基本的，它们只是从基本笔划组合出来的。实际上，我们可以用任何符号来代表复杂而常见的组合，例如：

口　代表　"3232"

但"口"只是组合，不是基本符号。它只是方便打字时用，在基本符号中不是一个新的符号，基本符号还是1,2,3,4,5,五个数字。正如在计算机的boolean algebra的0，1两元制。"100"等于十进制的"8"，"8"当然是比"100"容易写，但"8"并不是基本符号。

因此，我们在电脑打字时，为了方便，可以把许多常用组合以符号形式制订多划键，例如：

サ　233

シ　114

㇇　2314

以增加速度。但每一个新的符号都能还原于其基组符号1,2,3,4,5,这些都是组合符号，不是基本符号。

基本代码和为方便而设的组合代码是应该严格区分的。把两者混合，将会使代码丧失它的全面性和灵活性，而又只会增加它的局限性。最明显的就是影响到字典、图书索引的问题。

例如，打字中最常见的"口"作一个基本符号，首先把"口"固定下来后，便丧失了一些灵活性，现在把口＝3232，所以"弓"(2口31)里可用，"殳"(口245)和"巨"(2口2)里也可用。把"口"固定了，应用范围反而没有不固定下来方便。

其次把"口"固定下来，在查字典时就出现了"口"的特殊地位。例如，把它当作"6"，那么，口字部的索引都划分到"6"去，而其他同样的多组合边旁，如："シ"水字部、"木"木字部归纳到"1""2"开始的索引上，造成索引上的混乱。但归根结底是概念上的混乱，把"全部"和"局部"混淆起来的后果。

从逻辑上，当然也有一个问题，就是既然把"口"当作基本，为什么不把"シ"、"±"当作基本呢？但是，如果只把1,2,3,4,5,当作基本，则电脑打字机上的多划组合键，如"口""シ""±"……等，可多可少，可用可不用。随情况、随环境、随打字员的熟练程度而决定。

从上面几何、代数和逻辑的论证，"・一丨／＼"是最基本的组成部分，因最基本的是最简单的、最灵活的和最广泛的。这些组成部分的长短和写法都是不重要的，最重要的只是

把字里的任何笔划归纳到这几个类别去，而在电脑打字过程中把这笔划的程序按类别的代码根据中文写字笔顺输入电脑。

2、 基本笔划键盘设计

(1)引言

基本笔划方案是汉字编码中的一个非常简单的方案。只用1,2,3,4,5,五个数字来代表五种不同的笔划“·，一，丨，／，＼”。从原理上来说，基本笔划的电脑打字机只需用五个键子就能把所有汉字打出来。但一般手提打字机都可以容纳较多的键子，例如英文打字机一般就有四十个以上的键。因此，除五个基本键以外，可增加一些多划键，使熟练的打字员能打得更快、更方便些。增加什么样的键和怎样安排它们的键盘位置，是本文的主要内容。

通过科学分析设计键子和键盘，我们希望能达到“四个一”即，‘,,’一分钟内了解方案的原理，一小时内能初步试打，一天内完全学会，一星期内全部熟练。

键盘的规格和普通英文打字机相类似。见图（一）。因此这部中文电脑打字机还具有处理英文的功能，所以需要一个英文键盘。手指的基本位置与英文打字机完全一样。因此，一个懂中文的熟练英文打字员，可以在一个星期内学会，并高速度地打写中文。还有，在键盘上并能用上所有的键，即：除基本键外还能自如地应用多划键，那么每个汉字的平均频率为3.67键次。比英文每个字平均五个键次还要低，甚至也比现今流行的电报汉字代码的四个数字短。键盘上一共有五个基本键和三十一个带有代码的多划键。一个不会英文打字的中文打字员学习全面应用键盘的时间和英文打字差不多。两个星期的训练便能全部触打。因为每个字的平均键次低，所以效果要比英文打字好，速度应该比英文打字高25%。

和英文打字机比较，中文电脑打字机还有一个好处，就是一开始学便能马上触打，因为五个基本键便可以构成所有的字。所以，一开始便可以用基本键触打。随着熟练程度的提高，打字员可以增加利用多划键（其中有14个已成单字）。不象英文打字机必须把26个字母的位置完全牢记后才能触打。

以下我们谈谈怎样科学地设计键盘。

(2) 键的选择

让我们首先考虑一个常见的汉字：

“哲”－它的编码是231444233232一共十二个数字。

如果我们的键盘只有五个基本键：1,2,3,4,5,那末打字员就打十二键次。假设K是打的键次，即K＝12。

如果键盘可以容纳多划键，打的键次便可以减少。。例如包括两划键（一个键代表两个数字）。㉔键按一下便代表2、3两个编码：“十”或“下”或“冂”。从理论上说，两划键的形式是

xy　x＝1,2,3,4,5

　　y＝1,2,3,4,5

每个数字（x或y）有五个可能，两个数字便有25个可能：

xy＝11,12,13,14,15

　　21,22,23,24,25

……

51,52,53,54,55

如果要包括所有两划的可能性，便要添25个键子。

有了这25个键，上面的“哲”字便只需要打6下便成。所用的键是：

㉓　⑭　㊹　㉓　㉜　㉜

键盘的键子数由5个增加到30个，便可以将每个汉字的每字平均键次K减少一半。以五个键子的键盘打汉字，如果平均是每个K＝12，30个键子的键盘，便可成为K＝6。

如果我们的键盘再增大一些，每个汉字的键次还可以下降。例如，加3划键，如231代表编号2，3，1，的“十”或“”。从理论上来说，三划键的形式是：

xyz　　x＝1，2，3，4，5

y＝1，2，3，4，5

z＝1，2，3，4，5

那末，总的键数是5^3＝125个键。“哲”字便只用打四下：

231　444　233　232

130个的大键盘K＝4，比起30个键盘的K＝6，又少了两下。如此类推，如果我们再加大键盘容纳5^4＝625个键，每个键代表四个数字的编码，“哲”只需要打三下，就是K＝3：

2314　4423　3232

键盘最大的极限，就是一个汉字一个键，每次只打一下。那就与现在的大键盘无异。

总括来说，键盘小，学习容易，但每个汉字打的次数较多；键盘大，学习困难，但每个汉字打的次数较少。

表面上键盘的大小与打字的次数是对立的。但通过大量的材料统计，我们的中文电脑打字机最后的键盘设计是一个效率高的小键盘。上述的对立面是通过分析汉字的特殊性解决的。

汉字有两个特点：（一）有些笔划组合出现频率特别高，如“口”（3232），“扌”(2314)……等，而另一些如“戈”虽然很熟悉，但频率却很低。我们的键盘，只选择频率的较高组合。**（二）一些笔划组合比其他的容易辨别，因而比较易记易学。例如“十”（23）和“人”（45），就比“厂”（44）较易辨别。在出现频率相差不大时，我们挑选那些比较容易辨别和容易接受的组合作多划键。

汉字出现的频率很特殊，和英文比较它更集中。在五十万字统计中，常见的二千字占97%；而在这二千字中，头一百个字占45%；在这一百个字中，头十个字占12%。而单一个字“的”就占所有字出现的3%。在图（二）我们把每字的出现频率和字的出现多寡次序画出来。通过这些统计，我们选用了下列8个字编到键盘上：

有、的、口、是、我、不、女、日(1)

有些是因为频率很高，如“的”；有的是因为它们还可以当多划键。而当多划键时频率

** 统计频率的部分资料，是来自新加坡南洋大学华语研究中心(1976)。许乐斯，陈光泽，卢绍昌根据当地新闻所用约五十万字的文章作了统计，从所得数据编写的“新加坡华语常用字研究”。我们在这里把部分研究成果综合地讨论一下。并附有图表。

较高或用时比较复杂，如："不"、"女"、"日"、"口"。

在二划键中，例如24"厂"，42""等，我们将5×5＝25个所有可能的二划键和它们在约五十万字中出现的频率绘出来，表示在图（三）中。我们可以看出，二划的组合在汉字中也是非常集中的。我们挑选了五个频率高的编到键盘上。

14ノ　　43亻　　(2)
23十　　45人
24厂　　总计5个

在选编键码的过程中，有一个数据比频率更关键的是功用数u。功用数u就是打字员在打几十万字中，用这些新的多划组合键所省的键次。例如用"23"一次就省打了一下，用"231"每次就省二下。在表（一），我们把每一个二划键独立用时所省的实际数目列出。如果把功用数u，除去总的打字数，就是平均每汉字所能省下的数J。例如，加多一个键"23"，就能省了＝1.29下，而加一个键"25"就能省J＝0.015下了，相差一百倍之多。（详细的数学定义见附录）。

在三划的组合中，我们也把它们每个独立用时的功用数u和平均每汉字所省下数J都算出来，我们在图（四）把结果绘出来，同样地，最先几个组合占有大的功用数，以后便降低得很快。我们便挑出一些大u的，和易分辨的。它们是：

114　㇀　　(3)
121　𠂉
231　丁
232　土
233　艹
245　大
321　乚
323　㇂
421　𠂇
424　𠂆　　总计10个

在超过三划的组合中，我们把在"汉字基本构件频度表"中的250个部首及各种组合的独立功用数"u"，省下数"J"，都计算出来，选了大"u"的和易辨别的：

1231　讠　　(4)
124245辶
2314　扌
2324　尸
2341　木
3231　川
2322　丰
422231钅　　总计8个

上面四种键（1）－（4），便构成了我们图（一）中键盘上的31个多划键。

在选编键码时，我们是把每一个笔划组合的“u”和“J”独立地算出来的，但当我们把几个混合起来一起用时，就有一个新的问题出现。例如：有了“231＝㇆”的组合，就不能用“23”的组合。反过来有了“23”的组合，就不能用“231”的组合。因此，各个组合的功用数“u”，平均省下数“J”，都互相影响。在采用各个不同笔划组合时，以前各个独立计算得的功用数和省下数“u、J”都会改变得更少。我们用计算机把这些情况考虑进去。一群不同笔划组合键就等于一个新的键盘，每改变其中的一个，都要重新一起再计算。通过大型电脑计算，每计算一次就等于打字员打几十万字所得出的结果。

经过很多尝试才得出图（一）的结果。他们的功用数“u”和省下数“J”如图(五)。他们累积的省下数“J”和每汉字频率平均的键次数“K”，都在图（六）。在图(六)中我们看出：在曲线K越往键数多nK走的时候，它越改变得少。例如有nK＝20键时，平均键次K≒4.5下，而再加十一个键也不过减到K≒3.7。

在编选键码时，我们把人的因素也考虑进去。一些笔划组合虽然功用数“u”很高，但不“象样”，不易辨别的我们就不包括在内。

图（一）的36键小键盘虽然多划键不多，但都是通过统计证明是效率高才选编的。所以，平均每个字的键次可降为K＝3.67，当然，再多一些键，平均键次会再降低一些。但降低率越来越慢，不大值得。7个键的键盘，每汉字的平均键次K＝7.1，十个键键盘K＝6，而21个键，K可减为4.2。如把键数增到49个，K也只能减为3.4左右。

按照英文打字机的经验，键盘的键子数如果在36个上下，两三个月便能学会触打，如键数增多，触打便会困难。因此，中文电脑打字机键子的编选也以36个为限。经过严格的统计和打字员的实践，最后定案的多划键是效率最高和最方便的。

(3)键子的位置

把多划键编选了以后，跟着来的问题是键在键盘上的位置怎样安排才能使打字员感到最方便。从英文打字机的经验，我们可以归纳为两条：

（一）每个手指的灵敏度不同，以食指为最灵敏，尾指最迟钝。为着方便计算，我们给每只手指控制的键一个数1：

食指　1＝5

中指　1＝4

无名指1＝3

尾指　1＝2

（二）在触打时，手停的一排键(即基本及还原位置)是最易打，手停排之上的一排较难，下一排又再难一些，而最上一排则最难。现在给每排键一个数“K”：

容易排　K＝5

较难排　K＝3

再难排　K＝2

最难排　K＝1

那末，每一键便有两个数“1”、“K”，把它们乘起来“1k”，便有一个数字表示触打的难易程度。这些数字详细列在图(七)。

我们把使用频率最高的键子放在1K最大的位置，以后顺序而下，一直到把使用频率最

低的放在最难打的地方，也就是说“1K”最小的位置。结果便是图（一）的键盘。

（四）结语：

基本笔划方案的键盘是以科学方法设计，以合理、效率高和使用方便为原则的。在设计过程中有很多较细致的理论和较复杂的方程式，这里且从略不谈。当中也走过不少弯路，也浪费过不少电脑时间，这里也不一一赘述。只补充一句，就是：同样的笔划方案可用到日本和朝鲜文字的键盘设计里。我们准备在短期内试制出日文和韩文的电脑打字机。

表 1　二划键的功用数

顺　序	二　划　键	I 功 用 数 * U	J
1.	23	572,937	1.29
2.	32	532,233	1.27
3.	24	237,209	0.565
4.	31	202,745	0.483
5.	22	199,470	0.475
6.	42	187,964	0.448
7.	12	181,642	0.432
8.	11	141,972	0.338
9.	45	130,033	0.310
10.	21	121,687	0.290
11.	14	97,310	0.232
12.	43	97,285	0.232
13.	13	92,468	0.220
14.	33	69,074	0.164
15.	41	56,794	0.135
16.	34	49,652	0.118
17.	51	33,331	0.079
18.	44	24,199	0.058
19.	52	23,394	0.056
20.	54	12,695	0.030
21.	53	10,467	0.030
22.	15	8,156	0.019
23.	25	6,481	0.015
24.	35	228	0.0
25.	55	162	0.0

* 独立算计

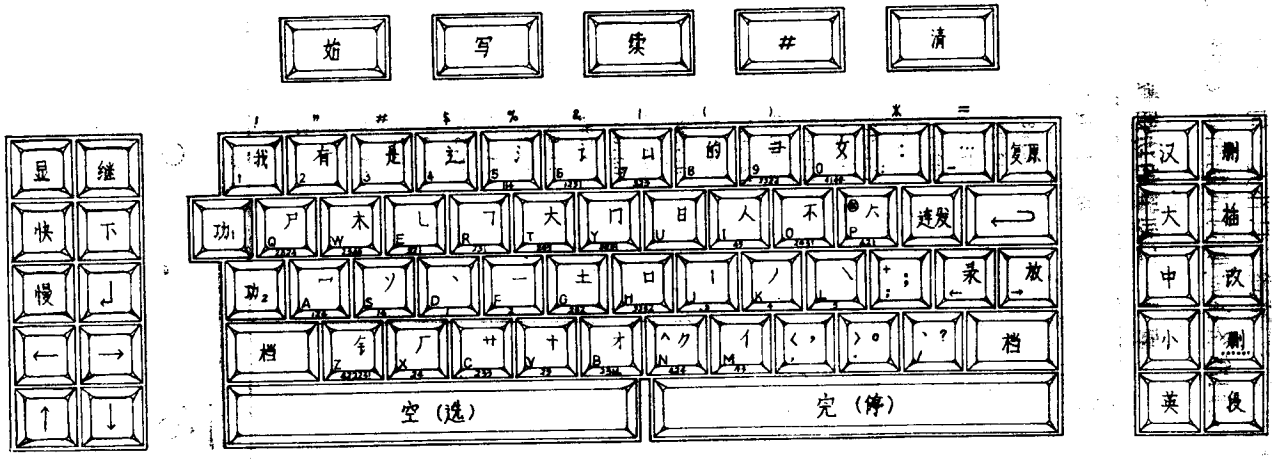

图 36个基本键和多划键的键盘

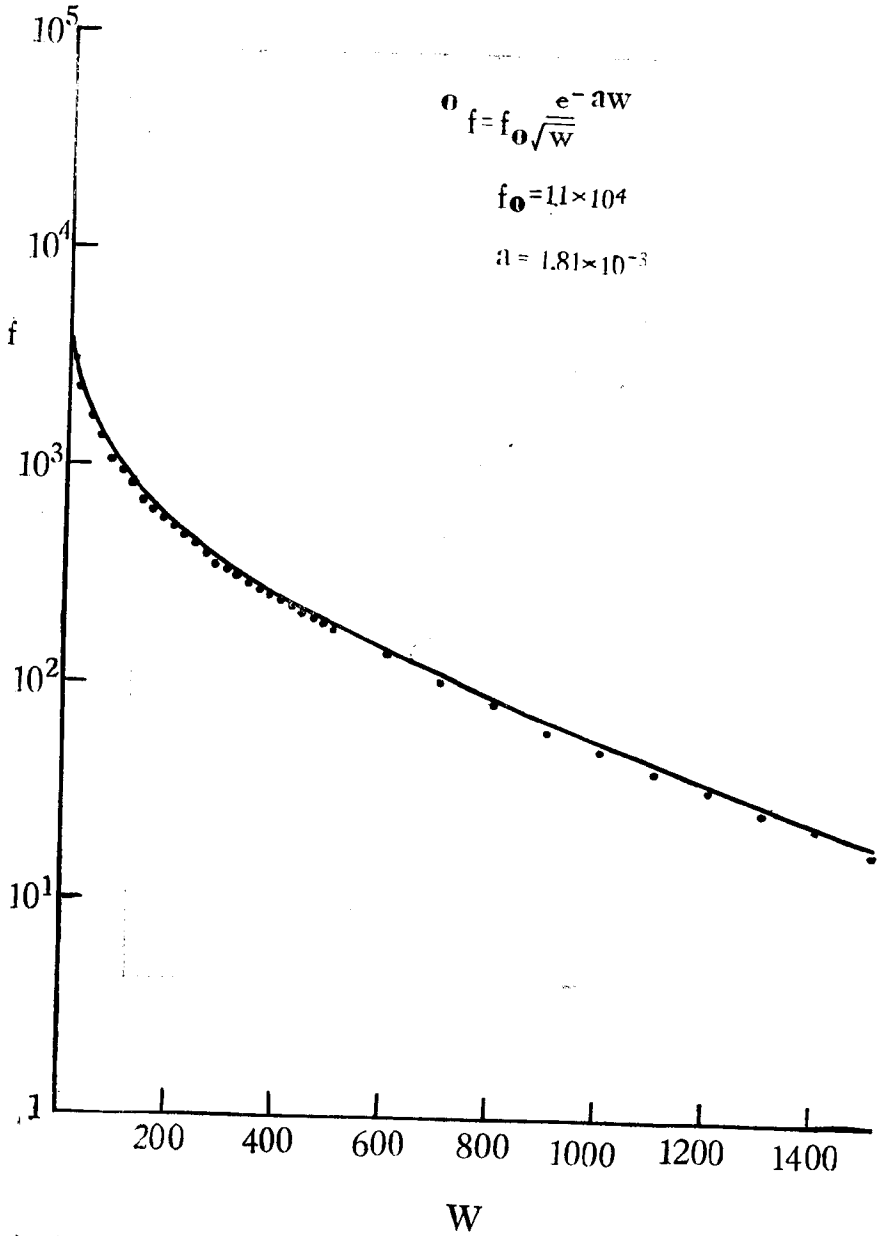

图2 汉字最常见2，000字出现的频率图。f 是每字出现的频率(在约五十万字中)。W是按频率高低而把2，000字排列起来的次序。他们可以约略地用$f_0 \cdot e^{-\alpha w}/\sqrt{w}$的方程代表，实线便是这个方程式

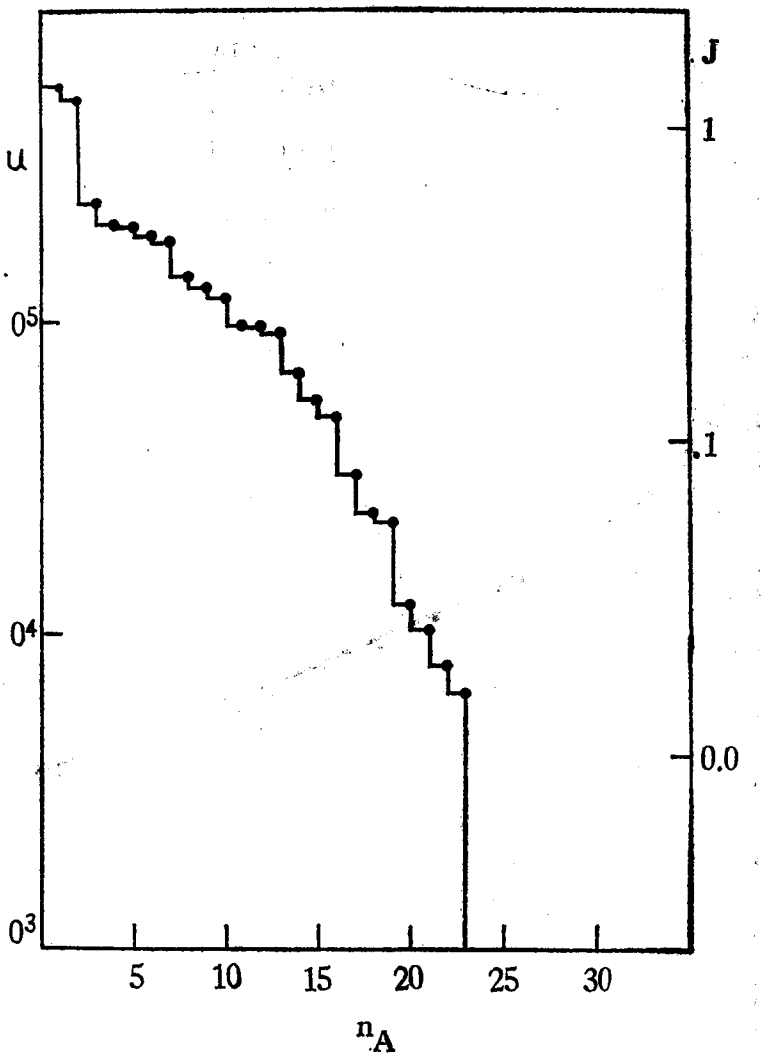

图3 划键的功用数u和省下数J，nA按功用数大小而排列起来的顺序nA＝1……25

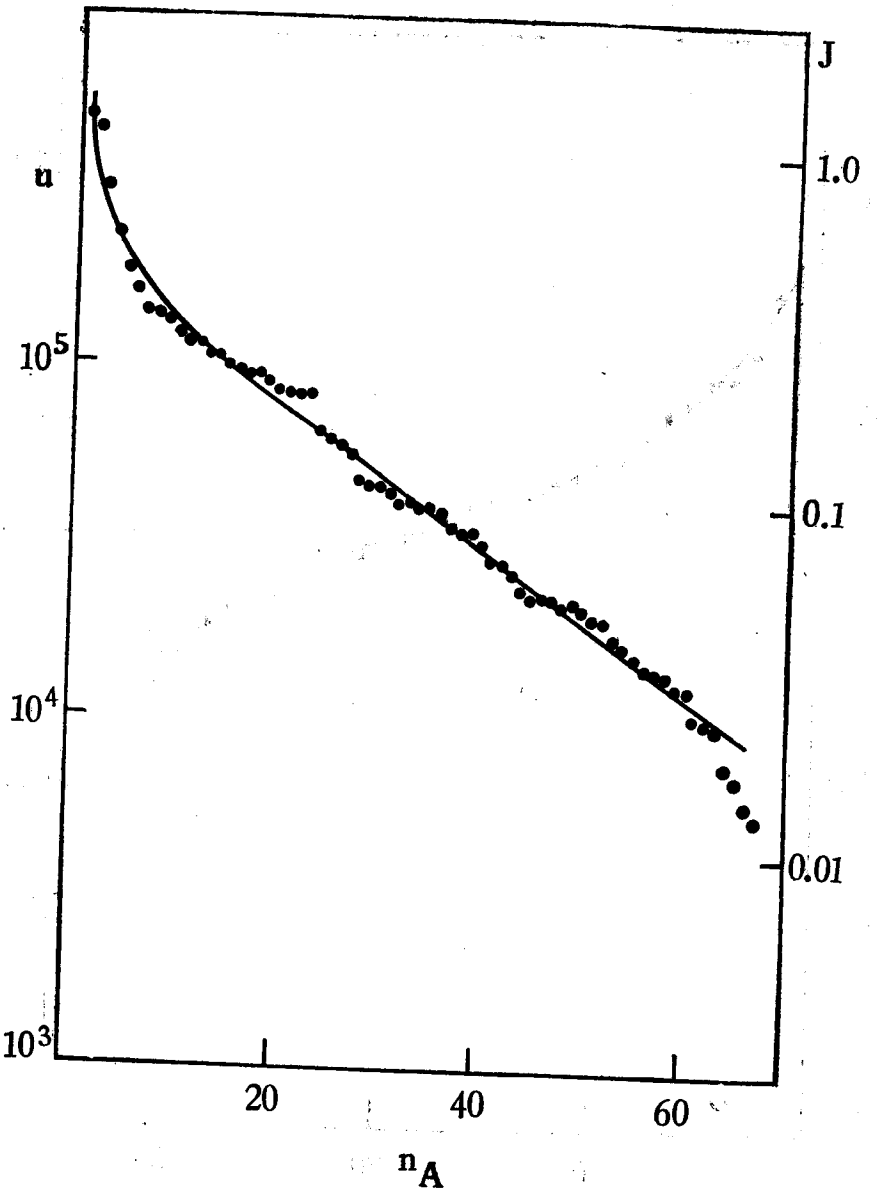

图4　三划键的功用数u和省下数J，nA是按功用数大小而排列起来的顺序

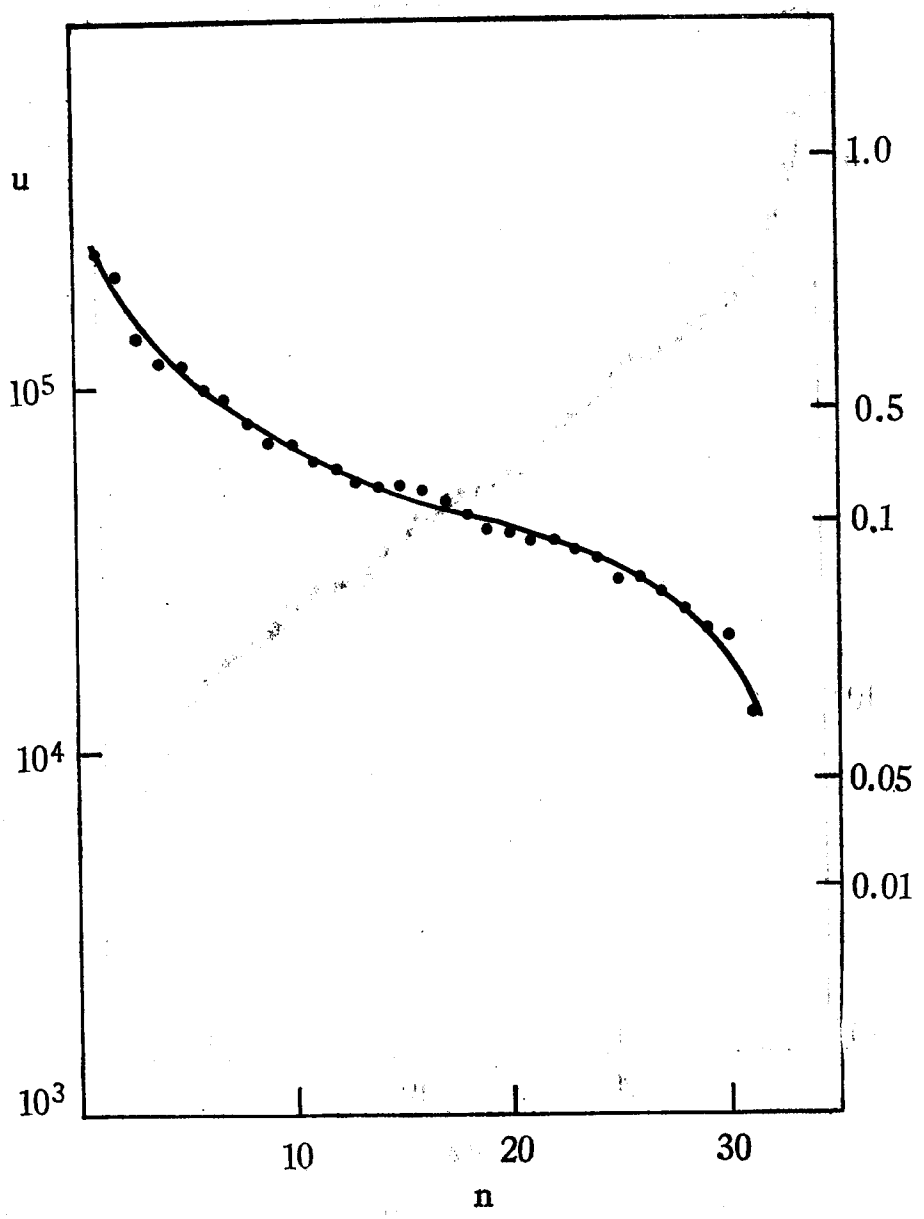

图5　31个多码键中各键的功用数u和省下数J。n是按功用数大小而排列起来的顺序。本图的结果是所有键的使用一起算，和图3、4每个键独立计算不同