# 技术参考篇

# TECHNICAL REFERENCE

目录



备注: ① 恺兴公司为康宝科技私人有限公司的前身

② 精简部首输入法 RIM (Radical Input Method) 原文译为 KIM

③ KIM (Knowledge-based Input Method) 中文原文译为恺兴输入法

④ “恺兴”中文书处理系统为“大学士”中文文书处理系统之先驱


康宝科技私人有限公司

BrushWriter Corp Pte Ltd

# 评价汉字输入方法的一种模式

王弗荃
(电脑工程博士，国立新加坡大学)
(Institute of Systems Science, National University of Singapore)
赵舜培
(康宝科技新私人有限公司)
(BrushWriter Corp Pte Ltd)
蔡恒胜
(加拿大卡尔顿大学中国访问学者)
(Dept. of Systems & Computer Engineering, Carleton University, Ottawa, Canada)

**摘要:**

中文电脑系统优劣的关键很大程度上取决于输入系统的性能，而汉字输入系统的性能又紧密地和键盘的设计、编码的规则及系统内部所具有的智能有关。此外，卖方提供的训练和用户本身的技能和经验也会影响输入效率的高低。本文将回顾一些商用的中文电脑系统、比较它们的输入方法，仔细研讨影响输入性能的因素，最后提出汉字输入方法评价的一种模式 —— “模糊模型”。

## 一、引言

近年来，汉字的编码输入技术得到迅速的发展，汉字的编码方案已多达五百余种，商品化的输入方法亦达几十种，随着研究的深入还会出现新的编码方案，如何对为数众多的汉字编码方法作出客观的评价已成为普遍关心的问题。

目前中国大陆和台湾已分别对各自现存系统的输入法进过一次评测，并公布了评测的结果。本文不打算对他们的评测方法做评价。只是根据我们的研究，讨论以下二种因素对汉字输入系统性能的影响: (1) 性能改进和提高的可扩充性; (2) 汉字输入的效率。

第一种因素是经常被人们所忽视，我们认为它是评价输入性能好坏的重要标志，一个好的系统必须具备为以后性能的改进和提高留有发展余地，也就是说该产品应具有较长的生命力。比如说要考虑汉字数量的进一步扩充，考虑既具有简体字又具有繁体字输入的功能以及可能增加常用词组的输入能力; 也要考虑如何进一步降低手指的疲劳度、减少重码率和每字的击键数等。今后所有这些改进应

该只在系统级上完成，而不应影响用户的使用。也就是说不应在编码体制上做大的改动，从而使用户需要重新学习一些规则。第二是输入效率，主是指把汉字输入到系统里的效率，可用来估价该系统可达到的最高输入速度。这里需要指出的是我们认为由卖方或发明者宣称的键入速度并不是个可靠的指标，因为它受人为因素的影响，比如操作员所受过训练时间的长短、操作员本身技能的高低以及对键入文本内容的熟悉程度等。实际上卖方提供的指标通常是一般用户无法达到的极限指标，而且操作者不可能在比较长的时间里保持这种键入速度。此外还有些系统通过击少量键就可提取予先存储或由用户定义的词组或短语，这无疑将大大加快键入速度，而卖方宣称的键入速度往往又以这种速度为依据，但如果找不到所需的词组和短语时，他们则必须逐个地键入字符，而这时输入的速度将比宣称的速度要低的多。鉴于这种观点，我们并不认为这种键入速度是衡量系统性能优劣的一种重要指标，而只考虑以上二种因素并认为它们是影响性能优劣的主要因素。

本文中我们将对几种实用的输入方法进行分析和比较，其中为了便于分析起见，对恺兴公司研究提出的 KIM 精简部首法考虑有两种执行过程，一种是系统含有词组输入，另一种是不含有词组输入。

目前在汉字输入方法上有几种评价的模式，在卢遂现提出的方案中(参见[1])，定量地进行了比较，他考虑了五种设计的因素，即键盘的尺寸、符号的数量，键入字符分解的步数、键入速度和检索的便利性。卢的方案揭示了输入方法的一个重要方面的度量，即学习和掌握的难易程度，但他假定各种性能的权值是相等的，而实际并不是这样。在中国大陆和台湾的一些资料 [2]、[3] 和 [4] 中讨论的方案涉及了广泛性能的评测，但事实上仍可能不是完美无缺的，也需要进一步完善。

本文在这里提出的模式只打算评价汉字的输入方法，而并不打算涉及执行这些方法的整个中文电脑系统。诸如汉字库字符的数量、硬件的性能等，仅看成是执行中的问题，并不在这里考虑。



## 二、几种实用汉字输入方法的比较

所有的汉字输入方法大致可归为三类: (1) 整字汉字识别法； (2) 以汉字发音特征编码为主的音符法； (3) 以汉字字形特征编码为主的形符法。

到目前为止，整字汉字识别法从技术上讲不如音符法和形符法成熟，而且在某些应用中，例如文字处理上并不一定有利。音符法的主要优点是简单，但为了得到它的输入码需进行转换，首先要把汉字转换成声音，然后按音符输入，因此不能盲打。为了有效地运用音符法还必须克服三个瓶颈: (1) 在发音上方言上的差异； (2) 发音较难的字； (3) 高的重码率。而相对讲形符法比较容易被人们所接受，这是因为汉字本身是象形文字，而且是独一无二的，形符法符合人们书写汉字的习惯。

在表 1 中列出了一些最常提及的实用汉字输入方法，包括恺兴公司的 KIM 精简部首输入法、王永民的五笔字形输入法、仓颉法、卢遂现的基本笔划法、李金铠的笔形码输入法、PC-CAI 计算机辅助输入法、支秉彝的见字识码、钱伟长的宏观字形输入法、大众拼形法、和 IBM 5550 键盘输入法。它们都是以形符法为基础，还有一些形符法没被列出是因为手头缺乏资料或被人们所接受还有待证明。

因此如果有一种汉字输入方法使得用户键入汉字如同他书写汉字一样的方便，不需要记忆任何硬性规定的编码规则，那来这种方法将最易被人们所接受，目前具有这类特点的输入法包括 KIM 法 [5]，卢法，PC-CAI 法，大众拼形法和 IBM 5550 的键盘输入法等。其中 KIM 法和 IBM 5550 的键盘输入法并不要求用户一定按正确的笔顺输入系统。

此外盲打的能力也是另一需要考虑的重要因素，如果一种输入法，每次输入都需要在屏幕上进行选择那末盲打几乎是不可能，输入速度必然是很低的。

## 三、汉字输入方法的性能测试

表 2 列出了以我们的模式对汉字输入方法性能评测的数值，也列出了这些测试中相对的权数。需要指出的是权数的赋值是随机的，它随着了潜在用户的需要而变化，这些用户既可能是专业打字员，也可能是偶而使用的用户。对专业打字员可以不强调学习和使用的难易，而更多强调减轻手指的负荷量和盲打的可能

性。但这些变化并不影响我们模式的精度。

下面我们来分析一下影响性能测式的因素：很多因素将会影响性能的测试，表 3 列出了影响汉字输入性能的主要因素，我们已经试图精确化有关性能测试的因素，并认为小的变化并不影响我们提出模式的精度。表中工程设计的因数包括编码规则、码的数量和选择，键盘并不影响我们提出模式的精度。表中工程设计的因数包括编码规则、码的数量和选择，键盘键数、键盘排列的质量和系统的智能。这里键盘的排列是指在键盘上码的位置的安排。它是影响每个手指负荷轻重的主要因素，合理的设计可使手指负荷均匀分布，效率最高。表 5 列出了手指和键盘的利用率，可以用来显示键盘排列的质量。系统的智能是指系统内用来帮助用户输入所具有的能力。例如精简部首输入法 (KIM) 可以在笔划级、字符级及词组级实现。在笔划级，KIM 使用群分析 (Cluster Analysis) 来处理检索字符时的错误，在字符级，它使用了向后链锁技术 (Backward Chaining) 用来来重码字进行智能选定，即根据前一个字和这个字的关连而自动选入合于逻辑关系的字。在 PC-CAI 法里也用了这种技术。KIM 还采用了向前链锁技术 (Forward Chaining)，所以在不等你输入完毕时，往往可把你所需要的字提前的字提前找到。在词组级，IBM 5550 笔触式的方法、王永民的五笔字形法、李金铠的笔形码输入法、大众拼形法和精简部首法等都配备了予先储存的词组和短语，即用定义的少数键来加快检索。

## 四、评价汉字输入效率的一种“模糊模型”

表 4 展示了每一种属性对汉字输入效率的影响，通过总结所有这些因素的影响可构成了一种“模糊模型”。在表 4 中可看出所涉及所有属性都是模糊组，也就是说判断这些属性是好是坏，它们是否可被接受并没有清晰的或轮廓明显的界限。然而通过总结每一种这类宽的模糊值，可以获得一个狭窄的模糊值，可用它来判断汉字输入方法的效率。

手指的负荷量、重码率、编码规则的数量、键数、在范围零至正无穷内码数的变化，这些值越小，这些属性就越好。

对于一种属性 X，它的作用函数可以用一负的指数函数来表示，即：

$$f(x) = w(x) * \exp(-x/x') \quad \text{..... (1)}$$



这里 w(x) 是在表中所给的权值 x，x' 是个常数，在 x = x(x 为 x 的平均值)

$$f(\bar{x}) = w(\bar{x})/2 \quad \text{.....(2)}$$

通过把 (2) 代入 (1) 式，可得到 x' 的值，(1) 变为：

$$f(x) = w(x) * \exp(-0.693 * x/\bar{x}) \quad \text{..... (3)}$$

表达式 (3) 可用来测量每一种属性的作用，如表 4 所示。这些属性的平均值可被计算如下：

手指移动的距离 (FTD)、交替打字的不可能性 (IAT) 和手指到达的困难程度 (DRF) 的平均值可从表 5 中得到，表 5 包括了几种输入方法的平均手指的利用率 (AFU) 和平均键盘的利用率 (AKU)。不同手指的相对权数的赋值是按照手指的灵活性、准确度以及按照手指是否容易达到为依据进行键盘的排列 (类似的原理也用于卢法 [1] )。为了简化起见，假定手指移动的距离从中心排移到最上一排、本排、下一排和最下一排分别为 +2，+1，0，+1 和 +2 个内部键距单位，那末 FTD 的值可通过键盘利用率和所有这些排移动的距离的乘积之和得到。FTD 平均值可以取列在表 5 中那些输入方法的 FTD 平均值计算获得，得出其值为 0.8065 个内部键距单位，IAT 是一种度量打字的负荷如何在左右手之间进行分配，可以取左手利用率和右手利用率的绝对差计算获得 IAT 的平均值为 11.33%

DRF 是一种度量手指到达不同排的困难程度。它可计算如下：

$$DRF = \underset{\text{手指}}{\text{SUM}}\{\text{手指利用率} * \text{手指权} * [\underset{\text{排}}{\text{SUM}}(\text{排利用率} * \text{排权数})]\} \quad \text{..... (4)}$$

这里 SUM 是求和函数，从表 5 DRF 平均值可算得为 3.44。

重码率 (DR) 值、编码规则数目 (#RULE)、键数 (#KEY) 和码数 (#CODE) 可取自表 1，它们的平均值分别为 40%，5，30 和 130。

盲打的可能性 (PTT)，系统智能的级别，键盘排列的质量 (KA)，训练和经验

的量 (TE) 不能用数字来表达，但是我们可以用下列原则赋一个二进制值 (0 或 1) 给这些属性:

(a) 对于 PPT 值，按照参考资料 [1]、[2] 在涉及大、中尺寸的键盘或者需要从键盘或屏幕连续选择时，对不允许盲打的那些输入方法 PPT 值应该置为 0；几乎所有以发音为基础的音符法都需要对同音字判别。因此它们的 PTT 值也为 0，除此之外，其它的输入方法都置为 1。

(b) 对于 SI 值，如果在字符级这种方法有智能，则 SI = 1/3；如果在词组级有智能，则 SI = 3/3；其它则 SI = 0。

(c) 对于 KA 值，对于键盘排列质量的简单判断不可以从手指的利用率和键盘的利用率中导出，在表 5 里可看出仓颉键盘并不好，因为手指的利用率并不随其灵活性和准确度而变化，而键盘利用率也不随不同排的可达性而改变，因此仓颉的 KA 值应该为 0，其它那些设为 1。

(d) 对于 TE 值，为了简化，我们可假定打字员已经得到了足够的训练，在打字前他们都具有同样的经验。在这种假定下对于各种输入方法 TE=1

表 4 总结了这种评价模式——模糊模型。

从上讨论可看出，模糊模型可以用来判断一种输入方法的输入效率。例如它可评估一种输入方法的键入速度，即:

S 评估{字符数/分} = Smax {击键数/分}*IE{%}/NKC{击键数/字符}

.....(5)

这里 Smax 是一个平均用户可达到的每分钟最大的击键数，即可达到的最大到速度，根据英文打字员的一般速度大约在 450 击键数/分；IE 是输入效率；NKC 是每个字符的击键数。若置 IE = 100%，则可算出一种输入方法可达到的上边界值。



## 五、对汉字输入方法较为完整的评价方法

除了输入效率外，在评价一种输入方法时还应考虑其它因素，应从工程的观点和应用的观点来考虑。

工程的观点是考虑有关系统设计的优劣，按重要性排列必须考虑以下因素: 编码规则的执行、键盘的排列、内部具有有智能、键盘的尺寸、码制的数量和选择。重要性的等级可从表 2 和表 3 里找出，表 6 展示了其计算值。为了介释表 6 中的项目，让我们先考虑编码规则，因为它对盲打是否可能将起主要的作用，赋权值为 3；其次，键盘的排列和尺寸也会影响盲打，因为它们的影响力是中等程度，均赋值为 2；这样编码规对盲打的相对影响为 3/7。编码规则对整个性能总的影响亦可算出，只要把每次测量的权数和编码规则对该次测量的相对影响的权值之乘积加在一起即可。

从应用的观点来考虑，按重要性排列的次序为输入效率、系统性能改进和提高的可扩性。

总之，在评价一种输入方法时以下问题应加以考虑:

从工程上看(按表 6 的重要性进行排列):

<1> 编码规则:

a. 一共有多少条编码规则? 都是什么?
b. 这些规则是始终如一的吗?
c. 是否可进行盲打?
d. 输入时需要思考、寻找吗? 需要花多大力量进行记忆?

<2> 键盘布置的设计:

e. 它是按照不同手指的灵活性和准确度以及键盘上的不同排手指的可达性来按排的吗?

<3> 系统的智能:

f. 系统能容忍输入错误吗? 有纠错功能吗? 有向前和向后链锁功能及自学习能力吗?

g. 有词组输入吗? 有上下文分析功能吗? 有多少词组量?

<4> 键盘尺寸:

h. 键盘的尺寸有多大?

<5> 码的数量及选择:

i. 码的数量有多少? 它们是根据科学研究进行选择的吗?

从应用上看:

<6> 输入效率:

j. 输入方法的效率是多少? (从表 4 可得到)

k. 估测可得到多大的键入速度? (从表达式 (5) 可得到)

<7> 改进及提高的可扩性:

l. 不改变码制规则和键盘设计可进行改进和提高吗?

## 六、结论

上面讨论了一种比较完全的汉字输入方法的评价过程，但大多数是关于定性的研究而不是定量的研究，对于输入效率的粗略表示，人们可使用表 4 这种模糊模型，并可用它来估算汉字系统的键入速度。根据我们的模型，可推算出 KIM 法的输入效率，在不包括词组输入时，大约为 0.69，估算的汉字键入速度为 80 字/分，这与我们实际的测试相似。在包括词组输入情况下，推算的 KIM 输入效率和键入速度分别为 0.74 和 112 字/分。对仓颉和天马(一种以发音为基础的音符法输入)系统，按我们的模式进行推算，其输入效率分别为 0.58 和 0.69，而它们的键入速度分别为 57 字/分和 63 字/分。我们现在展示的模型仍然是很粗略的，一些输入方法的资料也不完全和准确，当获得便多信息时，可以进一步进行改善。在本文中展示的观点只打算起个抛砖引玉的作用，引起人们关注，以便激发进行更多类似的研究，而文中的观点我们并不一定坚持全都是正确的，我们将接受正确的建议，随时准备进行修正。



☆图表说明:

表 1 — 某些以形符法为基础的实用输入法比较;
表 2 — 性能测试的权值赋值表;
表 3 — 影响汉字输入效率的因素表;
表 4 — 输入效率的模型;
表 5 — 手指负荷的度量;
表 6 — 按重要性排列效率的计算值。

☆名词说明:

KIM — 恺兴精简部首输入法;
Wang, Yongmin — 王永民的五笔字形输入法;
Cangji — 苍颉输入法;
Lo, SY — 卢遂现的基本笔划输入法;
Li, Jinkai — 李金铠的笔形输入法;
PC-CAI — PC—计算机辅助输入法;
Zhi, Bingui — 支秉彝的见字识码;
Qin Winchang — 钱伟长的宏观字形输八法;
Da Zhong — 大众拼形输入法。

**Table 1 - Comparison of Some Commercialized Shape-Based Chinese Input Methods.**

| Input Method | Coding Method & Remarks | # of Keys | # of Codes | # of Char. | Duplicate Rate | Keystrokes per Char. | Touch Typing | Claimed Speed (Char./min)vs. Learning Time | Commercial Systems |
|---|---|---|---|---|---|---|---|---|---|
| KIM | radical/character + Artificial Intelligence | 36 | 112 | 7000 | 2.57% | 2.9-3.4 | yes | 40 after 14 days<br>80 (skilled typist) | KTP<br>Macintosh Plus |
| | with phrases | | | | 1% | 3 | | 110 (skilled typist) | Atari 1040ST<br>IBM PC |
| Wang, Yongmin | radical + last-stroke code + shape code + 6 coding rules | 26 | 140 | 7000 | 0.02% | 2.6 | yes | | CCDOS on IBM PC/XT/AT |
| | with phrases | | | | | | | 160 (skilled typist) | Great Wall 0520 |
| Cangji | radical + 14 coding rules | 24 | 79 | 27255 | 4.73% | 4.59 | yes | 57 after 15 days | ChinaStar<br>IBM PC |
| Lo, SY | stroke + radical/character | 36 | 36 | 4000 | 5% | 3.7 | yes | 40 to 60 | CS-4000<br>APPLE II |
| Li, Jinkai | stroke + 3 coding rules | 8 | | 7000 | 1.3% | 3-6 | yes | 76 | IBM PC/XT/AT<br>BJS 130 |
| | with phrase | | | | | 2 | | | |
| CAI | stroke + radical +backward chaining | 34 | 88 | 7000 | 7% | 2.6 | yes | | BCM-286<br>IBM PC/AT |
| Zhi, Bingui | shape + sound + 10 coding rules | 36 | | 7000 | 5% | 1 to 4 | | 40 to 60 | IBM PC/XT |
| Qin, Weichang | radical + stroke/character + 13 rules | 39 | 145 | 7000 | 4.1% | 3 | yes | 20 after 14 days<br>50 after 21 days | IBM PC/XT/AT<br>IBM 5550 |
| | with phrases | | | | | | | | |
| IBM5550 keyboard | radical + stroke count + on-screen selection | 36 | 44 | 7000 | | | no | | IBM 5550 |
| Touch-pen board | radical<br>(2-character phrases provided) | 0 | 287 | | | | | | |
| Da Zhong | radical/character | 26 | | 7000 | | 3.42 | | | IBM PC/XT |
| | with phrases | | | | | | | 120 (skilled typist) | |

**Table 2 - Assignment of Weights to Performance Measures.**

| Performance Measure | Breakdown of Weight | | | Actual Weight |
|---|---|---|---|---|
| 1. Provisions for Improvements | | | 5% | 0.05 |
| 2. Input Efficiency | | | | |
| 2.1 Finger Load (Fatigue) | | | | |
| 2.1.1 Finger Travel Distance | 60% | 10% | 95% | 0.057 |
| 2.1.2 Probability of Alternate Typing | 10% | | | 0.0095 |
| 2.1.3 Difficulty of Reach of Finger | 30% | | | 0.0285 |
| 2.2 Duplication Rate | | 15% | | 0.1425 |
| 2.3 Keystrokes per Character | | 25% | | 0.2375 |
| 2.4 Possibility of Touch Typing | | 30% | | 0.285 |
| 2.5 Ease in Learning and Using | | | | |
| 2.5.1 Mental/Memorization Effort | 60% | 20% | | 0.114 |
| 2.5.2 Effort in Locating Keys | 40% | | | 0.076 |
| | | Total = | | 1.0000 |



Table 3 - Factors Affecting Efficiency of a Chinese Input Subsystem.

| Measure \ Factors | Engineering Design | | | | | Others |
|---|---|---|---|---|---|---|
| | Coding Rule(s) | Keyboard Layout: #of Keys | Keyboard Layout: #& Choice of Codes | Keyboard Layout: Arrangement | System Intelligence | Training & Experience |
| 1. Provisions for System Improvements | ● | ○ | ○ | | ● | |
| 2. Input Efficiency | | | | | | |
| 2.1 Load (Finger Fatigue) | | | | | | |
| 2.1.1 Finger Travel Distance | ○ | ◎ | ◎ | ● | | |
| 2.1.2 Probability of Alternate Typing | | | | ● | | |
| 2.1.3 Difficulty of Reach of Fingers | | | | ● | | |
| 2.2 Duplication Rate | ● | | ◎ | | ● | ○ |
| 2.3 Keystrokes per Character | ● | ○ | ● | | ● | ○ |
| 2.4 Possibility of Touch Typing | ● | ◎ | | ◎ | | |
| 2.5 Ease in Learning and Using | | | | | | |
| 2.5.1 Mental/Memorization Effort | ● | ○ | ● | | ● | ● |
| 2.5.2 Effort in Locating Keys | | ○ | ○ | ● | | ● |

Remarks:
- ● indicates strong influence, weight = 3
- ◎ indicates moderate influence, weight = 2
- ○ indicates weak influence, weight = 1

**Table 4 - Model for Input Efficiency**

| Input Efficiency Measure | | Contribution to Input Efficiency |
|---|---|---|
| 1. Load (Finger Fatigue) | | |
| 1.1 Finger Travel Distance | (FTD) | 0.057xEXP(-0.693xFTD/0.8065) |
| 1.2 Improbability of Alternate Typing | (IAT) | 0.0095xEXP(-0.693xPAT/11.33) |
| 1.3 Difficulty of Reach of Finger | (DRF) | 0.0285xEXP(-0.693xDRF/3.44) |
| 2. Duplication Rate | (DR) | 0.1425xEXP(-0.693xDR/4.0) |
| 3. Number of Keystrokes per Character | (NKC) | 0.2375xEXP(-.693xNKC/3.38) |
| 4. Possibility of Touch Typing | (PTT) | 0.285xPTT |
| 5. Ease in Learning and Using | | |
| 5.1 Mental/Memorization Effort | | |
| 5.1.1 Number of Coding Rules | (#Rule) | 0.02631xEXP(-0.693x#Rule/5) |
| 5.1.2 Number of Keys | (#Key) | 0.00877xEXP(-0.693x#Key/30) |
| 5.1.3 Number of Code | (#Code) | 0.02631xEXP(-0.693x#Code/130) |
| 5.1.4 System Intelligence | (SI) | 0.02631xSI |
| 5.1.5 Training & Experience | (TE) | 0.02631xTE |
| 5.2 Effort in Locating Keys | | |
| 5.2.1 Number of Keys | (#Key) | 0.0095xEXP(-0.693x#Key/30) |
| 5.2.2 Number of Codes | (#Code) | 0.0095xEXP(-0.693x#Code/130) |
| 5.2.3 Keyboard Arrangement | (KA) | 0.0285xKA |
| 5.2.4 Training & Experience | (TE) | 0.0285xTE |

**Table 5 - Measurement of Finger Load.**

| Utilization | Finger Utilization (%) | | | | | | | | Keyboard Utilization (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Left Hand | | | | Right Hand | | | | Top | Upper Home | | Lower | Bottom |
| weight / Method | Little 4 | Ring 3 | Middle 2 | Index 1 | Index 1 | Middle 2 | Ring 3 | Little 4 | 4 | 2 | 1 | 3 | 4 |
| KIM | 2.9 | 7.4 | 14.2 | 18.9 | 36.3 | 13.7 | 4.8 | 1.8 | 12.6 | 28.0 | 42.8 | 16.7 | 0 |
| Pinyin [Jain 85] | 5.5 | 9.5 | 11.0 | 24.0 | 26.0 | 10.5 | 7.5 | 6.0 | 12.5 | 28.0 | 42.0 | 17.0 | 0 |
| Cangji [Chen 86] | 6.48 | 5.22 | 6.16 | 19.01 | 38.62 | 9.57 | 11.14 | 3.80 | 0 | 40.1 | 35.4 | 24.5 | 0 |
| 3-point [Quo 85] | 1.56 | 4.59 | 5.82 | 35.1 | 39.73 | 6.13 | 3.58 | 3.47 | 10.11 | 18.67 | 24.00 | 15.23 | 31.98 |
| Dvorak [Jain 85] | 9.0 | 10.0 | 14.0 | 15.0 | 19.0 | 16.0 | 14.0 | 9.0 | | | | | |

**Table 6 - Calculation of Significance of Efficiency Measures.**

| Efficiency Measure \ Factors | Weight | Engineering Design Factors | | | | | Others | |
|---|---|---|---|---|---|---|---|---|
| | | Coding Rule(s) | Keyboard Design | | | System Intelligence | Training & Experience | |
| | | | # of Keys | # & Choice of Codes | Keyboard Arrangement | | | |
| 1. Provisions for Improvements | 0.05 | 3/8 | 1/8 | 1/8 | | 3/8 | | |
| 2. Typing Speed | | | | | | | | |
| 2.1 Load (Finger Fatigue) | | | | | | | | |
| 2.1.1 Finger Travel Distance | 0.057 | 2/9 | 2/9 | 2/9 | 3/9 | | | |
| 2.1.2 Probability of Alternate Typing | 0.0095 | | | | 3/3 | | | |
| 2.1.3 Difficulty of Reach of Finger | 0.0285 | | | | 3/3 | | | |
| 2.2 Duplication Rate | 0.1425 | 3/9 | | 2/9 | | 3/9 | 1/9 | |
| 2.3 Keystrokes per Character | 0.2375 | 3/11 | 1/11 | 3/11 | | 3/11 | 1/11 | |
| 2.4 Possibility of Touch Typing | 0.285 | 3/7 | 2/7 | | 2/7 | | | |
| 2.5 Ease in Learning and Using | | | | | | | | |
| 2.5.1 Mental/Memorization Effort | 0.114 | 3/13 | 1/13 | 3/13 | | 3/13 | 3/13 | |
| 2.5.2 Effort in Locating Keys | 0.076 | | 1/8 | 1/8 | 3/8 | | 3/8 | |
| Total = | 1.0000 | .2921 | .1527 | .1512 | .1669 | .1582 | .0922 | |



# 建立在知识层次上的智能型汉字输入系统

赵舜培，
（忆兴科技私人有限公司，Kaihin Technology PTE LTD Singapore)
王弗荃，
（电脑工程博士，国立新加坡大学
Institute of Systems Science, National University of Singapore)
蔡恒胜
（加拿大卡尔顿大学中国访问学者
Dept. of Systems & Computer Enginneering, Carleton University, Ottawa, Canada)

**摘要：**

本文描述了一种建立在知识层次上的智能型汉字文字处理的输入系统。这个系统是基于部首和字符的混合输入法。其独特之处在于该系统在笔画级、字符级和词组级具有智能。这样在部首、笔画串输入时，它能够自动识别和容忍非标准的笔顺和某种程度的笔形及笔画增漏的错误，从而找出用户所需要输入的汉字。这个系统又依据汉字和部首使用频率的统计规律，对常用汉字和部首进行了科学地优选，确定了用 36 个键（26 个英文字母键和 10 个数字键）代表精简的部首和 36 个常用字的输入方案，由于所描述的输入方法又同汉字的书写习惯及汉字字典的查阅方法基本相同，不需要特别记忆，因此很容易被具有一般中文基础的人们所接受。

## 一、引言

书写中文是由大量的表意文字 —— 汉字所组成。这些汉字是表示概念、物体或说话声音的符号，它是一种由象形文字发展起来的文字。构成一个汉字，它是基于几种基本的笔画以及由这些笔画构成的部首拼合而成，它们排列恰好在一方块内，所以汉字又叫方块字。由于它的衍生向来以形为本，没有拼音语系“字母”的观念，因此具有四千多年历史的中国文字还没能够完全解开以拼音文字为基础发展起来的现代计算机的全部潜力。

经过十几年的努力，在汉字的存储和输出技术上发展迅速，汉字的输入尽管提出了几百种方案，但至今还没有那一种方法为国内外所公认，从输入汉字的方法来分可以分成大键盘整字输入和标准小键盘即英文键盘的编码输入，而后者基本上又可分为两类：一类是根据汉字的发音进行编码，即音符法，另一类是根据汉字的字形进行编码，即形符法。由这二类基本的编码方法又派生出了字音和字形相结合的编码方案。

音符法主要的困难在于方言上的差异，对于不会普通话或方言较重的人使用较难。另外五万个上下的汉字仅用一千二百多个音节音调的组合来代表，则必然引起字符和它们发音之间的多字同音问题，因而必须辨别由此而造成的大量重码问题。此外，在输入每个汉字时，由于存在两步转换即先由字符转声音，再由声音转为音符，进行间接输入，所以不能期望输入速度太快。

而形符法却没有以上这些问题，因为每个汉字都是独一无二的。而且大多数中国人都熟悉使用笔画和部首，例如需要查阅字典或受过书法训练。因此形符法很容易被人们所接受。形符法按照笔画或部首作为基本单元又可分为笔画编码法和部首编码法。但无论是笔画编码还是部首编码，用户必须知道怎样把汉字正确地分解成笔画或部首，然后按照正确的书写顺序即笔顺，把它们输入计算机。现存的编码方案例如李金铠的笔形码、王永民的五笔字形码和仓颉码等方案，都要求用户根据一定的规则进行拆字，然后按照正确的顺序输入到计算机里，因此用户必须熟记这些规则。这对一般人来讲就变成一种负担，而且人们写字往往有自己习惯的写法。事实上不可能有所谓绝对标准的笔顺，人们写一些字常常会漏失或添加一些笔画或部首，甚至写错字，所谓的画蛇添足是不可避免的。但现存的输入方案都不能解决这类问题。因此探索一种有科学依据，简单方便，基本上不需要记忆，输入速度快，又有识别和检错功能的输入法仍是必要的。下面要讨论的恺兴输入法就是一种建立在知识层次上的智能型输入法。



## 二、KIM 的科学原理

从上讨论我们可再继续引深得出以下三点：

第一、应该说整字输入是最方便和最直观的，所需按键次数少，单字输入速度也最快，但由于汉字数量繁多，即使常用字也上千，这就给键盘的设计和排列造成困难，而且即使是最佳设计的大键盘，要从密密麻麻的几千字中找出所需要的汉字也是非常难和费时间的。

第二、从部首输入看在数量上比整字要少得多，按照“说文解字”归类共 540 个部首，（“说文解字”，东汉许慎撰，是我国语言史上第一部分析字形，说解字义，辨识声读的字典），随着古汉语的进化和发展，到了清代，著名的康熙字典部首体系已压缩为 214 个。现代的字典，虽然不同的字典有增有减，但总保持在 200 个部首左右。（例如，辞海 250 个，新华辞典 189 个，现代汉语词典 188 个等）。但即使这样，对计算机输入而言仍嫌太多。幸好汉字的这些部首有两个显著特点：

(1) 适当地精选少量的部首，或更确切的说是字根，可以复盖大量的常用字 [3,20]。

(2) 复杂的部首总是由少量简单的部首所构成，因此可以根据对汉字基本构件的频度统计规律进行分析，优选恰好的部首数目作为基本部首，用于计算机的输入。

第三、为了对汉字的字形特征进行分析，必须从汉字的基本组成单元着手，而部首虽然可以构成汉字，但一方面太多，另一方面也不能全面代表所有汉字，存在着不完全性和不唯一性。即使一些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。因此认为部首是汉字的基本组成单元是不完全和不确切的。那末什么是汉字的基本组成单元呢?

从几何学的观点看，点和线是一切文字最基本的组成部分，而且对汉字讲，它完全可由最简单的直线所构成。很多人的研究都认为汉字最基本的组成部分是一些基本笔画组 [4,5,6,7,9,10,11]。随基本笔画的数量而异，有几种笔画编码法。例如 Lee [4] 曾提出四种笔画，Caldwell [5] 提出过 21 种笔画，Wong [6] 提出 10 种笔画，而 Chou [7] 则提出 6 种，王永民提出 5 种笔画，而李金铠则提出 8 种等。我们的研究提出不《丶一丨丿乀》五个基本的笔画组 [12,13]。这也是根据中国古代著名的书法家提出的笔画规则，如颜鲁公八法，张旭八法，卫夫人七条，欧阳询八法等。由于构成一个汉字需要很多笔画，它比部首编码法的速度要慢得多，而且用户必须知道怎样把字正确地分解成笔画，然后按正确的书写顺序即正确的笔顺把它们输入计算机。但是汉字没有绝对标准的笔顺，因此对于一种特殊的字形输入法是否容易使用，主要取决于用户对实际书写顺序的熟悉程度。根据以上观点我们提出了恺兴输入法 (KIM) 或又称精简部首法的汉字输入方案。这是在根据对汉字字频和基本构件频度统计规律分析的基础上提出的一种字符部首混合输入法，并应用了人工智能技术，使其具有初步识别文字和自动检错的能力。

KIM 采用了 36 键容纳精简的基本部首和 36 个常用汉字的键盘输入法，它恰好完全利用了英文键盘上 26 个字母键加上 10 个阿拉伯数字键。选择 36 个键是有一定依据的，这是因为键盘的键数与汉字每字平均键数存在着对立关系，键数少，学习操作容易，但每个汉字打的次数较多。反之，键数多，学习困难，但每个汉字打的次数较少，通过大量的材料统计和英文打字的经验，选定 36 个键是效率较高，能较好地介决键数和打字次数的矛盾 [13,15,27]。

KIM 根据汉字部首的特点，从康熙字典 214 个部首里优选了其中 36 个作为基本部首，而且根据笔画代码相同，可以合并在同一键上的原则，实际上 36 键可代表 77 个字根（一些是不完整的部首，所以称为字根）。因此在数量上足以拼合中华人民共和国国家标准字符集里的 6763 个常用汉字。选择这 36 个部首的过程分二步。首先从 214 个部首里选择最常用的 100 个 [20]，然后分析它们，再选择其中 36 个，主要依据：a) 它们出现的频度 b) 它们是否也同时代表



一个完整的字 [12]。

KIM 又根据汉字字频的特点，尽管各家统计，颇有出入，但总的可看出有两条规律：一是常用字或高频字的集中性，在频率表的头二千字占 97%，头一千字占 94%，而头一百字占 45%，在这一百字的头 10 个字占 12% [21]。KIM 从头一百字里优选了 36 个常用字，据统计这些字出现的频率占 25% [13]。二是常用字的生命力，组词能力特别强。KIM 选择 36 个常用字的过程也分二步，一是找出前 100 个最常使用的汉字 [19,21]；二是再从这 100 个里面优选 36 个，按照 a) 它们使用的频度，b) 用上面提及的 36 个基本部首构成这些字的复杂程度及组词的频繁度。实践表明 KIM 系统选择了这 36 个常用字后，使每个汉字平均键入的次数降为 2.9 到 3.1 次 [13]，这样大大提高了输入速度。

KIM 经过优选常用汉字和基本部首以后，又对键盘的排列进行了科学的研究 [16,17,19,21,22,24]，最后确定如图一所示 KIM 键盘布置图。


图一、 KIM 键盘布置

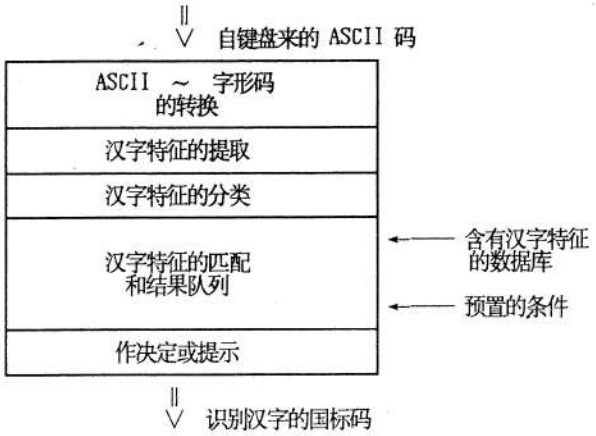
KIM 键盘有三个键用来区分基本部首，常用汉字和 ASCII 字符:

— 汉字／英文 (ENTER) 键: 允许用户在汉字和 ASCII 码间进行转换;
— 汉字结束 (RETURN) 键: 表示最后输入是一个常用汉字;
— 部首结束 (SPACE) 键: 表示最后输入是一个部首串。

从键盘上用户可直接输入用于中文文字处理的一些标点符号; 选 (TAB) 键的作用，在中高智能级 (下面要详谈) 时，当系统不能成功匹配时，可用选键来请求提示表。

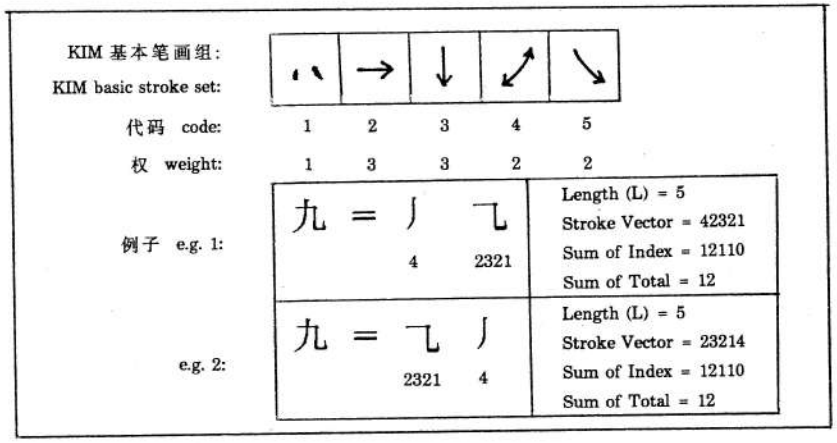
## 三、 KIM 汉字识别的 AI 解释程序

KIM 系统一个新颖之处在于应用了人工智能技术，使系统能够自动检查和识别在输入笔画或部首串时的笔顺以及笔形或笔画增漏的错误，并根据用户指定的智能级，还能容忍某种程度的输入错误，显示用户需要输入的正确汉字。 系统的这种智能是建立在包含有汉字字形特征的庞大的知识库基础之上的，图二展示了所应用的 AI (人工智能) 技术的梗概。 图三解释了 KIM 系统所使用的五种基本笔画及赋于的代码和权值。




图二、KIM 汉字识别 AI 解释程序






图三、 KIM 基本笔画组和代码、权的赋值

下面解释一下 KIM 汉字识别人工智能解释程序的流程简况:

### ASCII —— 字形码的转换:

这一步是首先对输入部首串里所包含的笔画向量化，以得到它们相应的字形码。 这样得到的笔画数称为这个字符的 N 值，笔画的顺序队列称作笔画向量 (Stroke Vector)，显然一个汉字如果有几种不同的书写方式的话，就能有好几个笔画向量。 图三里五种基本笔画组里的四种笔画都被向量化，剩下的一种或为点或为短笔画。

### 汉字特征的提取:

计算输入汉字的五种基本笔画的频率数，按次序排列的频率队列我们称之为索引和 (Sum of Index) 冠以字母 SI 表示。

汉字特征的分类:

每种笔画类型被予置为一个人为设定的权，例如“横” (一) 和“竖” (丨) 给于最重的权，因为它们具有最可靠的特征；而“点”或“短笔画”给于最轻的权，由于它的含糊性。 权和笔画的频率乘积之和称为总和 (Sum of Total)，以 ST 代表，它实际反映了一个汉字的笔形和笔画总的特征。

汉字特征的相配及结果队列:

将中华人民共和国国家标准 GB2312-80 《信息交换用汉字编码字符集》 基本集的 6763 个汉字 [24] 按照它们的 (N, ST) 对对应它们最普通的实际的书写顺序 (正确的笔顺) 进行预排列、对于一个汉字的输入部首串，有一个 (N', ST') 对，它的笔画向量将和地址为 (N', ST') 那个汉字的笔画向量相比较，如果它是唯一正确的能相匹配，则被匹配的汉字将显示，AI 的解释程序将着手进行下面一个输入部首串。 如果没能正确匹配或者有多个值相符合时，则系统将建立一个范围从 (N' ± 1, ST - I') 到 (N' ± 1, ST + I") 的比较区，这里 I'、I" 是用户指定的智能级赋于的值。 系统有三个智能级，即初级，中级和高级，它们的值可以是任意的，但在智能级和系统输入速度之间存有一种权衡。 在比较区里那些汉字的笔画向量 (S1 S2 S3 ..... Sn) 将同输入汉字的部首串 (S1' S2' S3'..... Sn) 相比较，以便搜寻最相近的匹配，这一过程称之为群分析 (Cluster Analysis)。

作决定或提示:

在不匹配或多个值相符合的情况下，AI，解释程序将提示一个汉字字符表，它们是按照队列 (Ranking) 值进行排列的，由用户从中进行选择，在正确匹配时，只显示所要求的汉字。 在实际执行过程中，KIM 提供了三个智能级。 “初级”在输八的部首串里最多允许有四个错误 (任意四个值) ，如果没找到所要的



汉字，则在屏幕上出现一个提示汉字表，供用户选择，“中级”允许有两个错误，“高级”只允有一个错误，这两级屏幕了动显示提示汉字表，除非由用户请求按选 (TAB) 键。 这二级是为了使那些有经验的打字员，实现更高的打字速度而设置，在所有的三级里，对输入顺序即笔顺的错误数并没有限止。

## 四、 KIM 系统的评

现有的中文键盘评价方案 [8,25,26] 并没能窥视 KIM 的全部潜力。 例如由魏悴 (WEI CUI) 提出的 14 条准则 [8] 和卢逐现提出的 7 条准则 [26] 都没有提及恺兴所具有的错误恢复能力。 而应用拉开菜单和老鼠定标器使得恺兴系统比其他系统对用户更友好，这是 KIM 系统的另一重要特点。

KIM 系统性能总结如下:

键数 = 36
符号数 = 51
每个汉字所需的平均击键次数 = 2.9 - 3.0 次
专业打字员单字输入的平打字速度 = 60 - 80 汉字 / 分 * ※。

---

* 平均讲每个汉字要比一个英文字包含更多的信息。

※ 这里指的打字速度只是在提供笔画级智能情况下单个字的输入，系统在字符级和词组级提供的智能将会大大提高输入速度。

表 1. 使用 WEI CUI 方案评价 KIM 键盘

| 序号 | 准则 | 权数 | 得分 |
|---|---|---|---|
| 1. | 编码符号的易识别性 | 10 | 10 |
| 2. | 不发音 | 10 | 10 |
| 3. | 编码符号 = 键符号? | 12 | 8 |
| 4. | 编码／键符号的比率 | 8 | 5 |
| 5. | 单级编码 | 5 | 5 |
| 6. | 编码规则的一致性 | 5 | 5 |
| 7. | 键盘尺寸 | 10 | 10 |
| 8. | 盲打条件 | 3 | 3 |
| 9. | 击键数 | 7 | 7 |
| 10. | 重码 | 10 | 3 |
| 11. | 字数量 | 3 | 3 |
| 12. | 码的可读性 | 3 | 3 |
| 13. | 成本 | 12 | 12 |
| 14. | 紧凑性 | 2 | 2 |
| | 合计 | 100 | 86 |
| | 规范值 | 1 | 0.86 |



表 2. 使用 LO 法评价 KIM 键盘 [26]

| 序号 | 变量 | 英文键盘 | KIM 键盘 |
|---|---|---|---|
| 1. | 键盘尺寸: 初学者 (L1) | 1 | 1.11 |
| | 专家 (L2) | 1 | 1.56 |
| 2. | 符号 (S) | 1 | 2.17 |
| 3. | 在打字时分析步骤 (P) | 1 | 0.67 |
| 4. | 在打字时困难的等级 (V) | 1 | 4.50 |
| 5. | 数据库索引系统: | | |
| | 存储器的作用 (I1) | 1 | 1.53 |
| | 搜索的作用 (I2) | 1 | 0.67 |
| | 合计 | 7 | 12.21 |
| | 规范值 | 1 | 1.74 |

## 五、 结束语:

恺兴智能型输入法已和恺兴中文文字处理软件包构成了一种非常实用的中文文字处理系统，后者具有窗口功能、卷帘式菜单和老鼠定标器辅助选择等特点，并能打印简体繁体、宋体、楷体和英俄德法日希腊西班牙等不同的字体和文种。目前，这种中文系统已经在好几种最普通使用的微机上执行，如 Macintosh Plus、Atari 1040ST 以及带有恺兴存储／图形板的 IBM PC, XT, AT 等。这种输入法应用人工智能技术在汉字形符输入上克服了两大困难：即系统可以容忍非标准的笔顺输入错和常见的笔画增漏及笔形的错误，具有了初步的识别文字和自动纠错的能力。进行的实验和获得的数据表明 [2, 12, 14, 16, 18]，只要用户能够写出的汉字，他就能输入这个汉字，而且不要求一定按照标准的书写顺序和绝对

正确的笔形输入。

该系统除提供在笔画级的智能化外，还在字符级及词组级提供智能，它采用了向前和向后链锁技术 (Forward and Backward Chaining)、自学习 (Self-Learning)、声音回馈 (Voice Feedback) 以及预存词组 (Prestorage Phrase) 等设计思想，使得汉字前后连接的逻辑性和组词能力强的特点得到了发挥，这大大提高了输入速度。系统还可根据用户的环境动态地更新数据库和扩大知识级的层次，并相应改进硬件的性能，以达到更高的智能级别。



**参考文献:**

[1] Chiu, A. (1986) U.S. Patent pending.

[2] "Kaihin Chinese Keyboard: Theory and Specification" Kaihin Research Technical Report RP-86-005, 1986.

[3] Suen, C.Y. & Huang, E.M., "Computational Analysis of the Structural Compositions of Frequently Used Chinese Characters" Washington, 1984, pp163-176.

[4] Lee, H.C. U.S. Patent No. 4, 462, 703.

[5] Caldwell (1960) U.S. Patent No. 2, 950,800.

[6] Wong, W.S. (1985) "Method for Encoding Ideographic Characters" U.S, Patent No. 4, 505, 602.

[7] Chou, H.C. (1979) "Input System for Sino-Computer". U.S. Patent No. 4, 173, 753.

[8] Wei Cui "Evaluaton of Chinese Character Keyboards" COMPUTER Magazine, January 1985, Special issue on Chinese Computing Systems, pp54-59.

[9] Zwi B. & Lo S.Y., "A Simple Stroke Ordering Code and Its Analysis" School of PHysics, University of Melbourne, Technical Report UM-P-83/6, pp120-132.

[10] Leung et al. (1983) "Means for Encoding Ideographic Characters" U.S. Patent No. 4, 379, 288.

[11] Wang G., (1982) "Chinese Printing System" U.S. Patent No. 4, 327, 421.

[12] "Radical Frequency Analysis of the 7,000 Chinese Characters" Kaihin Research Technical Report RP-86-002, 1986.

[13] "Experiments on the KIM Keyboard Layout" Kaihin Research Technical Report RP-86-003, 1986.

[14] "An Analysis of Commonly-Used Chinese Phrases for Computer-Assisted Chinese Word Processing" Kaihin Research Technical Report RP-86-002, 1986.

[15] "Kaihin Input Method: Summary and Examples" Kaihin Research User Manual UM-86-008, 1986.

[16] “Conversion Table for Chinese Language”, Kaihin Research Technical Report RP-86-008, 1986.

[17] Olson & Jasinski, “Keyboard Efficiency”, February 1986, pp241-244.

[18] “Chinese Character Data base”, Kaihin Research Technical Report RP-86-009, 1986.

[19] Suen, C.Y., “Computational Studies of Most Frequent Chinese Words and Sounds” World Scientific Publishing Co., Singapore, 1986.

[20] 郭平欣等编，“汉字信息处理技术”，中国国防工业出版社，一九八五年。

[21] 许乐斯，“新加坡华语常用字研究”，新加坡南洋大学华语研究中心，一九七六年，第 1 - 30 页。

[22] 现代汉语资料分题选编（上册），山东教育出版社，一九八三年。

[23] 信息交换用汉字编码字符集基本集 (GB 2310-80)，中华人民共和国一九八一年。

[24] 中文电脑输入法与输入器调查评估总结报告，财团法人资讯工业策进会，1984, 1, 台北，中国。

[25] Lo, S.Y. “A Scientific Model for Comparing Various Methods of Inputting Chinese Characters into Computer”, Computer Processing of Chinese and Oriental Language, May 1985, Vol. 2, No. 1, pp36-58.

[26] 卢遂现“一个中文打字机的草案”抖擞 1，一九七四年一月，第 19-27 页。

摘自《1987 年中文信息处理国际会议论文集 I》



# 人工智能在中文电脑中的应用

* 蔡恒胜 ** 赵舜培

* 加拿大卡尔顿大学系统和计算机工程系中国访问学者

** 加拿大悦兴科技研究中心主任兼总工程师

**摘要:**

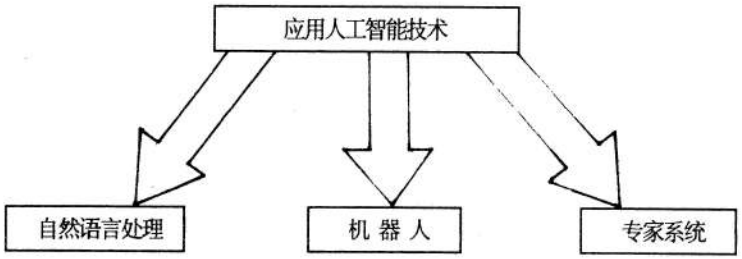
近几年来，人工智能的研究一直十分活跃，而人工智能和中文电脑的结合，正在解决以前中文电脑不能够解决的课题。本文将以精简部首中文输入法和中文文字处理软件为基础来讨论人工智能在中文电脑中的应用，着重探讨如何根据汉字的属性来建立适宜人工智能应用的环境，以及讨论应用人工智能技术后，在改进汉字键盘输入方面所能起的重要作用；最后展望人工智能将对中文电脑的发展产生的深远影响。

## 一、引言

自第二次世界大战以来，电脑科学家一直在试图发展一些技术以便使电脑能更像人的行为，例如决策系统、机器人、语音系统、视觉系统和各种专家系统等，这些通称为人工智能技术 (Artificial Intelligence — AI)；但目前大多数的努力仍停留在实验阶段。最近几年里，建立在知识层次上的专家系统 (Knowledge-based Expert System) 取得了引人注目的进展，越来越证明它们潜在的巨大实用价值。它使得电脑能模拟人类专家的行为来协助人们分析问题、进行判断和作决策。例如管理复杂的规划、诊断疾病、找矿、设计复杂的硬件和寻找故障等 (1)。除了专家系统外，人工智能研究的范围还包括另外二个分支：一是机器人系统，特别是发展可视的和能触知的程序，使机器人在某一环境移动时能观察前方的变化。另一主要研究重点是自然语言处理 (Nature Language Processing)。它包括自然语言的理解 (Language Understanding)、语音系统 (Voice System)、视觉系统 (Vision System)、学习与推理 (Learning and Inductive Inference)、以及用来表达以上各系统知识库的知识表示法 (Knowledge Representations) 所使用的工具 —— “人工智能程序设计语言” (AI Programming Languages)。图一展示了人工智能的概况。

近年来，将人工智能应用在中文电脑上也取得了一定的进展，但其意义不仅仅是这些，更重要的在于将解开作为与拼音文字相对应的表意文字 — “汉字”在使用电脑上的巨大潜力。本文首先将分析汉字的属性和汉语的特点，分析在信

息处理中，中文电脑和人工智能的结合正在揭示在某些方面是拼音文字所无法比拟的这种语言的优越性。随后讨论如何建立汉字特征的知识库和人工智能的解释程序；最后展望应用人工智能技术后，将对中文电脑的发展产生深远的影响。




图一、 应用人工智能的概况

## 二、汉字的特点及信息处理

中国文字通常指的是汉字，因为汉语和汉字是中国百分之九十以上人民的语言文字。汉字是从象形文字发展出来的，汉族最初创造的文字，已无实物可考。但可估计，黄帝以后，已经陆续出现了汉族的文字了。有实物可考的最早的汉字是 1880 年在河南安阳小屯村（殷墟）发掘出来的刻对文字的龟甲骨片，称为甲骨文，它是公元前 1300-1028 年间的产物，上距黄帝时代约一千多年，下距现在约三千多年。汉字向来以形为本，由形知义，是一种表意文字。虽然难学、难记、难写，但却是最具有表达能力的文字，它能表达精密的思想和感情，是记录语言的理想工具。几千年来我们祖国的伟大科学家、哲学家、文学家、政治家等的光辉成就，都是依靠汉字保存下来的。我们国家是世界四大文明古国之一，这也是与汉字能记载几千年来的丰富遗产分不开的。汉字经久不衰，具对顽强的生命力，它必然有着某些独到之处和特点。

我们从信息处理的观点来看，汉字具有以下优点:

1. 汉字是一种象形文字或者说是一种图形文字，在人的感官中，视觉是最敏感的，对概念的形成最直接了当。而图形对视觉则是最理想的识别符号，一张图等于千言万语。因此，以象形文字为依据的汉字远较借音转换的拼音文字产生的概念表达的强烈。



2. 汉字字形固定，在单位面积上表示的信息最多也最完整。目前全世界没有任何文字在单位面积上的变化有汉字多，而且采用完全相同的单位面积（例如英文长短不齐，且只有 26 个字母的排列变化）。利用这种特点可提高信息处理的效率。

3. 在概念的表达上，汉字用字最少。用字少表示在大量信息处理时，速度快，效率高，而且经济。许多报告中认为中文用字较英文要约少 20%-30% 左右。因此，同一份文件，在效率和经济性上，理论上中文比英文要高 50%-60%。

4. 汉字的突出优点是单音词特别多，组词能力特强。只要两个单音汉字，颠来倒去便能组成绝大多数的词汇；至于三个四个单音汉字，其组词能力就更复杂、更灵活。汉字的这种性能使汉字成为世界上极其难能可贵而异采独放的一种文字。(据统计在新华字典中，所收 8500 个字中，除开“囫囵”、“葡萄”等不能拆开的连绵词 (共 578 个，占 6.8%) 外，单音成义的单字竟达 7922 个字，占 93.2% 之多) (2)。

5. 汉语语音的利用将大大提高信息处理的效能。因为汉字为单音节，每字一音，比英文的复音节省，而且语音的变化很少影响文字。从古到今，不论何时何地，不论读音如何，字形永远受到认可。

另外虽然我们的语言不断吸收外来的词汇和语法，但汉语的基本词汇和句法是很少改变的。这对于建立汉字的知识库是十分有利的。但是汉字的突出难点是在于单字字量冗多，字形复杂，笔画繁多，这给信息处理造成了困难。简化汉字和汉字标准化以及根据汉字字频、词频和汉字构成成分的统计，实行分级处理和专业处理是十分重要的。

近年来，中文电脑发展很快，已有很多商用系统面世，例如长城系统、国乔系统、中文之星、天马系统等，它们多运行在 IBM PC 机上。很遗憾由于受硬件条件的局限，特别是内存容量小、速度慢和 CPU 的非线性寻址等缺点，使开发出来的中文软件也受到局限，尤其在汉字的输入上，方案繁多，各有千秋，但没有一种方法为人们所公认。因此我们比较了各家的方案，发现有几点是趋于一致

的: (1) 目前都采用和英文键盘相兼容的标准小键盘，进行编码输入；(2) 针对不同的对象，系统采用了同时共存多种输入法，由用户进行选择；(3) 一般多以形符输入为主，拼音为辅；(4) 为了提高输入速度，减少重码，又要使用户能掌握，都制定了一些需要记忆的规则；虽然有的系统也具有联想式等的智能特点，但受条件限制，应用有限。我们认为各种方案主要的问题在于输入的方法和人们书写汉字的习惯不同，需要另行一套规则，这对一般人讲将成为一种负担。然而如果按人们写字的方式编码输入的话则存在以下几个问题: 一是汉字的笔画繁多，通常在九画到十三画频率为最高，即使按部首笔画拼合计算其构件数平均也4-5 以上，换句话说由于输入码的长度长，需要击键的次数多，因而十分繁琐；二是人们写字往往有自己习惯写法，而且常常会漏失或添加一些笔画，甚至写错字，因此即使按标准笔画编码，也需要解决容忍一般人写字常犯的毛病。三是始终存在重码问题，如何进行限止和消除。而精简部首中文输入法和恺兴中文文字处理软件企图应用人工智能技术和其它措施来克服以上问题，使得输入汉字同汉字的书写顺序及查阅字典的方法基本相同，不需要特别记忆，因此很容易被具有一般中文基础的人们所接受 (3) (4)。

首先，精简部首中文输入法和恺兴中文文字处理软件是在基于准 32 位的 MC 68000 的 MACINTOSH PLUS 上开发的，该机容量大，速度快，特别是具有线性寻址和图形功能强等特点，因此为中文电脑和人工智能的发用创造了有利的环境。进而欲根据汉字的基本结构单元和汉字的属性为依据，在适宜应用人工智能技术的环境下，企图建立一个包含有汉字特征的庞大的知识库。然后利用群分析 (Cluster Analysis) 的三级检索、向前和向后链锁技术 (Forward Chaining and Backward Chaining)、自学习 (Self-Learning)、声音回馈 (Voice Feedback) 以及预存词组 (Prestorage Phrase) 等设计思想，使系统在笔画级、字符级和词组级具对智能。这样在输入笔画和部首串时，它能够根据用户指定的智能级，迅速进行检索和寻找所对应输入部首串的汉字，而且能自动识别和容忍非标准的笔顺和某种程度的笔形及笔画增漏的错误，从而显示用户所需的汉字。而恺兴中文文字处理软件能运行在升级的 Macintosh、MAC PLUS、Atari1040 以及带有恺兴存储/图形板的 IBM PCXT, AT 等机型上，它具有窗口功能、卷帘式菜单和老鼠定标器辅助选择等特点，以及能打印简体繁体，宋体楷体和英俄法德希腊等不同的字体和文种，使得这种输入法及软件成为一种非常实用和对用户极为方便的中文文字处理的工具。下面将以精简部首中文输入法和中文文字处理软件为基础来讨论人工智能在中文电脑中的应用及其效果。



## 三、精简部首输入法和汉字特征的知识库

精简部首输入方法是基于部首和字符的混合输入法，它采用了 36 键容纳精简的部首和 36 常用汉字，恰好完全利用了英文键盘上 26 个字母键加上 10 个阿拉伯数字键，能较好地解决键数和打字次数的矛盾。而 36 个常用汉字和精简部首 (或称字根) 是根据汉字字频和汉字构成成分统计规率进行科学优选的 (5) (6) (7) (8)。从几何学的观点看，点和线是一切文字最基本的组成单元，而且对汉字讲，它完全可由最简单的点和直线所构成。很多人的研究都归结为一些基本笔画组，随基本笔画的数量和形状而异，分别提出了不同的编码方法。我们的研究是按照《丶、一、丨、丿、㇏》这五种基本的笔画组，由这五种基本笔画构成了精选的基本部首(字根)，而由这些基本字根和笔画可拼合一切汉字。例如“花”字，我们可键入“艹、亻、丿、乚”的部首串，最后按“部首完键”即可。

我们对精选的结果进行了程序的验证，这个过程是这样的，对于五种基本笔画，首先将其笔画向量化，除点或短笔画 (丶) 外，四种笔画 (一、丨、丿、㇏) 都可向量化 (→、↓、↙、↘)，分别赋以码值《1, 2, 3, 4, 5》代表这五种不同笔形的笔画，而这五种基本笔画可排列成 P 种不同的字根，这里的 P 值为:

$$P = 5 \times 5 \times 5 \times 5 \times 5 + 5 \times 5 \times 5 \times 5 + 5 \times 5 \times 5 + 5 \times 5 + 5 = 3905$$

根据中华人民共和国国家标准 GB2312-80《信息交换用汉字编码字符集》基本集，我们将 6763 个汉字用这五种码值按其正确的笔顺进行编码，构成一个汉字的笔画库；然后分别统计它们的构成成份，即 P=3905 种不同的字根在这 6763 个汉字中组合的频率，优选高频的字根并考虑它们是否也同时代表一个完整的字及书写的难易程度，与实际精选的结果进行比较，完全是符合的 (9)。根据笔画代码相同，可以合并在同一键上的原则，实际上 36 键代表了 77 个字根。而为了减少每个汉字的平均击键数，我们又优选了 36 个最常用的汉字，单键输入，据统计这 36 个字在一般文章里出现的频率占 25%-30% 左右 (10)，有意识地使编码形式适应汉字出现频度不等的规律，使得有效按键数大为提高，在不包括词组输入时，汉字平均键入次数降为 2.9 到 3.1 次 (7)，大大提高了输入速度。

根据笔画构成汉字 (H) 的四个要素，即笔画形体 (笔形 γ)、笔画数量 (画数 N)、笔画的位置 (X) 和正确的书写顺序 (笔顺 S) 是影响汉字字形的四个变

量，可列出汉字的特征函数为:

$$F(H)=f(\gamma, S, N, X)$$

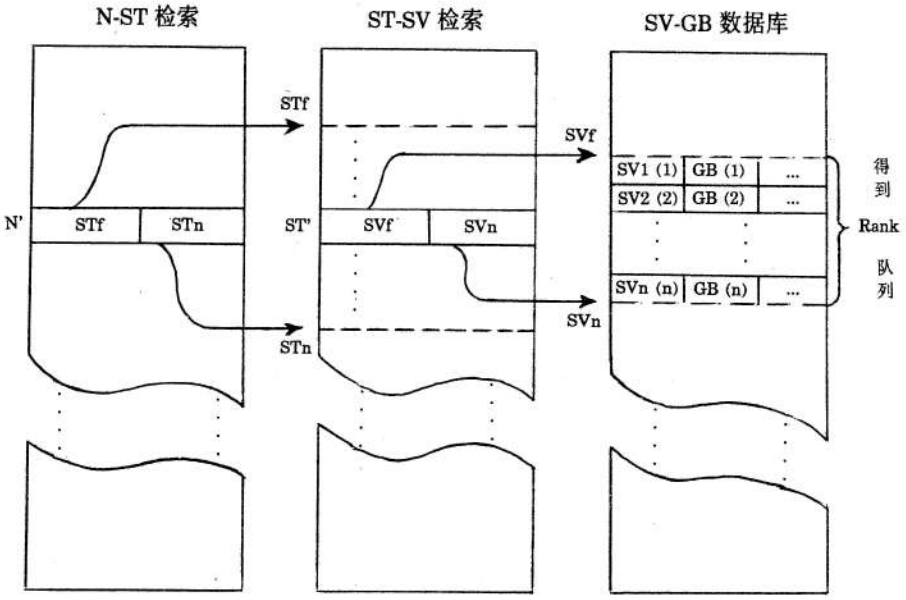
以这个汉字的特征函数为基础，我们从汉字构成的基本单元入手，来推导和建立包含汉字特征的数据库和人工智能解释程序。首先笔形变量 $\gamma$ 是依照我们规定的这五种基本笔画《丶、一、丨、丿、㇏》的笔形为依据，它们的码值分别为《 1, 2, 3, 4, 5 》；又按其稳定程度，即书写时相互是否容易混淆，预置一个人为设定的权值，分别为《 1, 3, 3, 2, 2, 》，这里横 (一) 和竖 (丨) 给于最重的权，因为它们具有最可靠的特征；而点或短笔画 (丶) 给于最轻的权，由于它的含糊性。权值间的差值实际表示了笔形之间差异的大小。对笔顺变量 (S) 我们用笔画码值的顺序队列来表示，称作笔画向量 SV，显然笔画向量 SV 和按其顺序书写的汉字相对应。但人们书写一个汉字笔顺不同，它的笔画向量就不一样。我们规穴一个用标准笔顺书写的汉字是对应它的笔画向量。暂不考虑笔画位置 (X) 这个变量，虽然会造成少量重码，但总共只造成 1701 多组的重码，可采用其它措施加以限制。笔画数 N 对某一汉字是个常量，它等于这五种笔画频率之总和。为了让系统能容忍某种程度的输入错误，再引入二个变量: 统计一个汉字的五种基本笔画的频率数，按《丶、一、丨、丿、㇏》的次序排列可得到一个频率数列，称为索引和 (Sum of Index)，冠以字母 SI 表示。把笔画的频率和它们权的乘积之和称为总和 (Sum of Total)，以 ST 代表，它实际反映了一个汉字的笔形和笔画总的特征。然后对于 GB2310-80 的 6763 个汉字，可建立起一个 (SV-GB) 数据库以及 (N-ST) 和 (ST-SV) 两个检索库 (11)。系统搜寻正确匹配或最相近匹配的三级检索过程我们称为群分析，这个群分析的程序是系统笔画级智能的 AI 解释程序，而系统的 N-ST 检索库、ST-SV 检索库和 SV-GB 数据库整个构成了包含有汉字笔画特征的知识库。图二展示了三级检索的过程。

此外，根据在现代汉语中，大多数汉字不是孤立出现的，它们一般是组成某种固定的文法结构。因此，汉字的出现概率是有条件的，即随前面已经出现的字的不同而具有不同的条件概率。例如对汉语词组的分析可发现以某一字为词头的词，多的有几十个，少的只有几个，还有为数不少的连绵词；除词组外我们还特别测试了在文本中字的前后逻辑连接频率，发现一个汉字后面经常跟着它出现的字并不多，这少数字占了在这种条件下汉字出现概率的很大部分，利用语言学中的这种性质，我们采用了称之为向后链锁 (Backward Chaining) 和向前链锁



(Forward Chaining) 以及自学习 (Self Learning) 的程序来提高编码效率，这是系统字符级的 AI 人工智能程序。系统还建立了可由用户设定和预先存储的词组或短语，这由系统词组级的 AI 人工智能程序所实现的。

在建立了这几级智能后，目前这个系统在汉字输入时，能容忍任意的非标准的笔顺的错误。例如“花”字，我们可键入“艹、亻、丿、乚”部首串，也可键入“艹、亻、乚、丿”，甚至还可键入“亻、丿、乚、艹”，都可迅速得到“花”字。系统也能容忍一定程度的笔形或画增漏的错误，例“花”字，如果键入“艹、亻、一、乚”或者“艹、亻、丿、丨、一”或“艹、亻、乚”，通常最相近的匹配也都可找到“花”字。按我们统计，通常在笔画向量上出现 4 个以下的常见错误时，经过最相近的匹配的搜索后，所要的汉字 90% 以上会落在提示行的第一组上，也就是只需多打一键即可 (12)。系统现在还配有三千多对词及词组，用户也可自建词组，这是十分方便的。我们利用向后链锁技术，当用户键入某组带有重码的部首笔画串时，即根据前一个字，系统可自动选择合乎逻辑关系的下一个字。例如一撇一捺可构成“人”、“八”、“人”三个重码字，但若前一字是“女”字，系统会自动选择“人”字；如前一字是“输”字则会自动选择“人”等等。又利用向前链锁技术，当用户在键入某个字后，例如“法”字，接下来要输入“院”字时，当他输入笔画“了、丨”时，系统就把最有可能输入的“院”字在提示行中显示出来，供用户选择，如不是用户所要的，只需继续键入其它键，直到找到他所需要的字为止。此外系统还具有“自学习”能力，以不太合逻辑性的单字组成词汇，如人名、地名、国名等，如“加拿大”是属于不合逻辑性的词组，当“加拿”这两个字被输入后，系统不能够在“大、又、丈”这组重笔字里选择“大”字，那是因为“加拿大”并不属于逻辑性的词组。不过电脑能够吸取第一次的经验，在第二次输入“加拿”两个字后，系统能自选择“大”字。系统加入这些功能后，不仅给操作员也给一般用户带来了极大的方便，同时也大大提高了输入速度。




图二、 笔画级 AI 人工智能程序的检索过程

**四、人工智能解释程序**

下面简单说明一下系统笔画级的 AI 人工智能解释序，对于一个汉字的输入部首串，系统首先获得它的笔画向量 (SV') 、由笔画向量 (SV') 可算出它的笔画数 (N') 和总和 (ST') 值，然后进行群分析，经过两级检索后，在 SV-GB 数据库里可得到一个系统指定的缩短的 SV 值的范围，将来个笔画向量值 (SV') 和范围里的 SV 的值相比较，如果它能唯一正确匹配 (exact match)，则匹配的汉字将被显示，AI 解释程序将着手进行下一个输入部首串。如果不匹配或有多个值相匹配时，根据设定的智能级，将建立一个从 (N' ± 1, ST+I" ) 的比较区。这里的 I', I" 的值是根据用户指定的智能级来确定的。系统设有三个智能级，即初级、中级和高级，它们是根据一般打字错误和打字率设置的不同的错误容忍值，使比较区的范围大小有所差异。然后系统将搜索最相近的匹配 (proximite match)，以获得一个 Rank 队列，并列出与输入部首串相近的那些汉字，即那些汉字的总和值 (ST) 与输入部首串的那个汉字的总值 (ST') 是十分相近的或相同的，而笔画数最大相差 ± 1 笔，供用户进行选择。

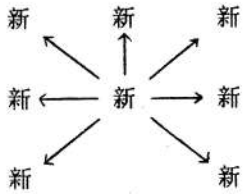


笔画级的 AI 人工智能程序，共进行三级检索和两步的匹配过程，从而得到所需的汉字。系统的硬件保证了有足够快的速度来执行这个过程。至于字符级和词组级的 AI 人工智能解释程序是根据字符库和词库查表检索，比较简单，故不再解释。

**五、系统的测试**

我们用计算机的程序对系统进行过大量的测试，验证了设计思想，得到了一些“量的概念”，对今后中文电脑的改进十分有益。

为了使系统的输入能够适应书写的习惯，我们在数据库里储存了一般人写字常犯的笔画上的毛病，下面我们来看一些例子，例如“新”字，通常一般人有这样八种写法:




分析部首串: 新 = 立 + 木 + 斤 = 丶 一 丶 丿 一, + 一 亅 丿 乀, + 丿 丿 一 丨

可发现对这些部首串人们经常的习惯写法不外乎有以下几种:

立、立 2 种;
木、木 2 种;
斤、斤 2 种。

总计一共有 : 2 x 2 x 2 = 8 种组合,

我们建立的笔画级的人工智能解释程序使得系统能容忍以上差异，因而不管你怎样输入，系统总可找到这个“新”字。而且其中三种“新、新、新”，通过群分析和匹配过程后，总排在 Rank 队列的第 0 列，即直接出现在屏幕的光标

位置上，其它五种“新、新、新、新、新”，则排在 Rank 队列的第 1 列，即在提示行的第一组的第一个字，也就是通过选字再多打一键即可。类似我们统计了 283 个汉字的字根的各常见的写法，获得相应的笔画向量 (13)，让数据库包含有 GB 码的 6737 个汉字常见的各种组合，因而使得系统能容忍常见的笔形和笔画增漏的错误。据程序统计 6768 个汉字，总的组合数为 164050 种，输入后没找到的组合仅为 483 种，占千分之二点玖 (0.29%)，其中落在光标位置有 5621 种，提示行第一组即 Rank 队列 1 - 6 列的为 561 种，这二种总计为 6182 种，占 91.3% (12)，程序统计的结果充分说明系统具有容忍绝大多数常见的笔形和笔画增漏的错误，并且与正确的部首串的输入相比，通常最多只需多打一键，至于非标准的笔顺错，人工智能解释程序的最相近匹配的过程本身就给以了保证，使系统能容忍任意的笔顺错。通过计算机程序的一系列测试，验证了我们的设计思想、数据库的设置和笔画级的人工智能解释程序的编制是相当成功的。

关于字符级的人工智能解释程序需要特别指出的是系统不仅仅依靠词的前后搭配，更重要的是依据字的前后逻辑关系来设定。我们也进行了程序的测定，发现有些常见关连字的前后搭配出现的频率是远高于一些词的搭配 (14)，因而我们字符级数据库的建立并不单纯是词库，而是包括关连字库。利用前后链锁和自学习程序使得系统对用户变得十分方便和友好。

## 六、结束语

以上描述了应用人工智能技术在汉字形符输入上克服了两大困难: 即系统容忍非标准的笔顺输入错和常见的笔画增漏笔形的错误，这是具有初步的识别文字和自动纠错的能力；而字符级的智能，系统利用了汉字具有逻辑连接和组词能力特强的特点，通过搜索建立的数据库，达到自动选择的目的。尽管目前应用人工智能还主要是在汉字的字形级上，也可扩展到字音和字义级的处理上，至于在汉语的语音、语意处理和句法、文法处理上，根据已得建立的从汉字的基本单元入手的应用人工智能的环境，今后可动态地升级和扩大知识级的层次和数据库，例如扩大自学习区和语义文法区，不断地扩大和增强 AI 人工智能解释程序的功能，并相应地改进硬件的性能，例如用多微处理器来提高检索的能力;采用超高速和超大容量的大规模集成电路来扩展内储容量和处理速度，进行新系统的设计等，以达到更高的智能级别。

目前人工智能的应用还在模式识别 (Pattern Recognition) 和语音识别 (Voice Recognition) 上，使中文电脑能够进行声音和印刷体或手写体的直接输入，例如最近光学字符识别 (OCR) 输入技术已取得很大进展，使得大批量字符信息的快速输入成为现实。至于在寻求解决各类汉语输入法 —— 如键盘输入、语音识别输入及光学字符识别输入等的障碍，可采用永存储器 (Permanent Storage) —— 电擦只读存储器 (Electrical Erasable ROM)，硬盘 (Hard Disk) 等去增强系统动态性的智能。总之，人工智能和中文电脑的结合有着广泛地发展前景，需要语言学家、社会学家和电脑科学家等共同努力，来解决以象形文字为基础的汉字在信息处理中的各种障碍。我们相信让电脑能理解汉语已并不是太遥远的事了。

参考文献:


(1) Harmon & King, "Expert Systems" pp1-pp12, John Wiley & Sons. Inc. 1985.

(2) 郭平欣，张淞芝"汉字信息处理技术"，国防工业出版社，pp26-38, 1985, 12.

(3) 赵舜培，王弗荃，蔡恒胜"建立在知识层次上的智能型汉字输入系统" Kaihin Research Inc. Canada ISBN: 0-921315-04-x.

(4) 王弗荃，赵舜培，蔡恒胜"评价汉字输入方法的一种模式" Kaihin Research Inc. Canada ISBN: 0-921315-05-8.

(5) "Kaihin Chinese Keyboard: Theory and Specification" Kaihin Research Technical Report RP-86-005, 1986.

(6) "Radical Frequency Analysis of the 7000 Chinese Characters" Kaihin Research Technical Report RP-86-002, 1986.

(7) "Experiments on the KIM Keyboard Layout" Kaihin Research Technical Report RP-86-003, 1986.

(8) "An Analysis of Commonly-Used Chinese Phrases for Computer-Assisted Chinese Word Processing" Kaihin Research Technical Report RP-86-004, 1986.

(9) "Radical Frequency Test Program of the 7000 Chinese Characters" Kaihin Research Test Report RT-86-001, 1986.

(10) "Test Report 1" Kaihin Research Test Report Rt-86-002, 1986.

(11) "Test Report 2" Kaihin Research Test Report Rt-86-003, 1986.

(12) "Test Report 3" Kaihin Research Test Report Rt-86-004, 1986.

(13) "Test Report 4" Kaihin Research Test Report Rt-86-005, 1986.

(14) "Test Report 5" Kaihin Research Test Report Rt-86-006, 1986.

(15) Hsu, L. S. "Artificial Intelligence and Chinese Computing", Proceedings International Conference on Chinese Computing, August, 1986. Singapore.