中 文 信 息 学 报
第8卷 第2期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.8 No.2

基于N联字的汉字识别后处理研究

苗兰芳 张 森 周昌乐

(杭州大学计算机系)

【摘要】为了提高汉字识别率，本文提出了在单个汉字的初级识别后，利用*N*联字的上下文关系，对初级识别中拒识或不确定的汉字语段作进一步确认的一种方法，阐明了*N*联字后处理方法的基本思想，给出了实现此方法的数据库的结构设计方案和理论算法，分析了理论上可提高的识别率，最后给出了一个*N*联字汉字识别后处理系统模型。

关键字: 汉字识别 *N*联字 数据库 后处理

一、前言

计算机汉字识别是指用计算机建立视觉系统，自动认识印刷或书写在纸上的汉字。随着模式识别技术和人工智能原理的日臻完善，汉字识别的研究得到了不断发展，近十年来，在印刷体，手写体汉字识别的研究领域内取得了一些可喜的成果，但当前存在的问题是识别率较低，还不能进入实用阶段，因此，如何提高汉字识别率，尤其是手写体汉字识别是计算机汉字信息处理领域中一个及其重要的课题。

从识别方法上来看，目前汉字识别方法的主流是统计方法和句法方法。这二种方法都已经有了较长的历史，积累了很多成熟的经验。但由于在实际应用中汉字书写技术的原因，仅仅使用上述方法已不可能提高汉字识别率[1]。因此，有必要另寻一条作为补充识别的方法。即在上述识别方法的基础上，对那些不能确认的汉字加以进一步处理，以提高识别率。这种对识别结果代码文件进一步加工，提高系统识别率称为识别后处理，简称后处理。

到目前为止，尽管有人提出根据汉字文本的上下文关系来进行后处理[2]，如：中文词切分后的完整匹配和自然语言理解上下文匹配，但都没有具体的实施方案。本文提出了一种基于N联词[3]（N为大于等于2的整数变量）的汉字识别后处理的方法，不但能较方便地实施，而且还能明显地提高汉字识别率。

①本文1993年5月17日收到

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、N 联字后处理方法的基本思想

为了能更好地表达问题，先作如下定义：

定义 1: 如果汉字 a_i (i＝1,2,3,…,n)具有前后联系，即能使 $a_1,a_2,a_3\cdots,a_n$ 有意义，则我们称 $a_1,a_2,a_3,\cdots a_n$ 为 N 联字，其中 N 为大于或等于 2 的整数变量，因此，当 N＝2 时，有二联字 a_1a_2；N＝3 时，有三联字 a_1,a_2,a_3。例如："人才、人称、大人、敌人"，是关于"人"的二联字；"主人翁、人生观、接班人、"是关于"人"的三联字。

定义 2: N 联字 $a_1,a_2,\cdots,a_n$ 也可称其为 a_i 在第 i 位的 N 联字。

例如：3 联字 a_1,a_2,a_3，也可称 a_1 第 1 位，a_2 第 2 位，a_3 第 3 位的 3 联字。定义 1 例子中，"人才、人称"是"人"为第一位的二联字，"大人、敌人"是"人"为第二位的二联字；"主人翁"是"人"为第二位的三联字。

定义 3: 联字频率：几个联在一起的字，共同出现的先验概率 f；

字频：某个汉字在所有汉字中出现的频率 W；

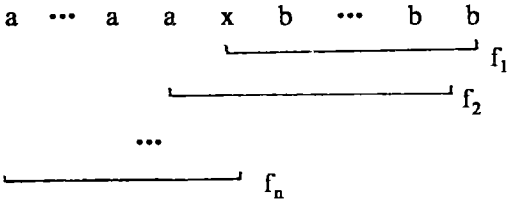
不确定汉字：在初次识别中不能确定的汉字，包括拒识汉字。

相似度 R：待识汉字和不确定汉字的相似程度。

在 N 联字汉字识别后处理法中，首先建立一个初级单个汉字识别中不确定的汉字库以及这些汉字在第 1 位第 2 位…直到第 n 位时的 N 联字字库及其相应的 N 联字频率，然后根据由单个汉字识别时所提供的不确定汉字的相似度 R，不确定字本身的字频 W 和不确定汉字在各位时的 N 联字频率 f_i (i＝1,2,3,…,n)的综合评判函数 G 值的大小来确定汉字。即：

$$G = G\ (R,\ W,\ f_1,\ f_2,\ f_3\ \cdots\ f_n)$$

式中$f_1,f_2,f_3,\cdots,f_n$为待识汉字在各位时的 N 联字频率；如图一所示：



图一

通常取综合评判函数为这些因素的加权和。即

$$G = K_0 \times R + K_0' \times W + \sum_{i=1}^{N} K_i \times f_i$$

式中 K 为加权系数。自然，采用 N 联字后处理法，在几个待确识的相近字中确定汉字时，某一个字的函数值愈大，识别为该不确定字的可能性也愈大。

具体地说，在汉字串$A = a_1,a_2,\cdots,a_i\ X\ a_{i+1}\cdots,a_m$中，x 为不确定汉字子集，即 $x=\{x_1,x_2,x_3,\cdots,x_s\}$(s 为子集字数)；$a_i$ (i＝1,2,3,…,m)为已识别汉字，那么，要确定 x 必需

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分别计算 x 对应每一个不确定汉字 x_j (j=1,2,3,…s)时的综合评判函数值，即要计算下列矩阵式:

$$\begin{bmatrix} G_1 \\ G_2 \\ \\ G_s \end{bmatrix} = K_0 \begin{bmatrix} R_1 \\ R_2 \\ R_s \end{bmatrix} + K_0' \begin{bmatrix} W_1 \\ W_2 \\ \\ W_s \end{bmatrix} + \sum_{i=1}^{N} K_i \begin{bmatrix} f_1 \\ f_2 \\ \\ f_{si} \end{bmatrix}$$

上式称为汉字识别矩阵式，$\{R_1,R_2,\cdots R_s\}$为单个汉字识别时给出的不确定子集中每个汉字和待识汉字之间的相似度，$\{W_1,W_2,\cdots,W_s\}$对应不确定子集中每个汉字的字频，它们由前人统计数字库给出，$\{f_{1i},f_{2i},\cdots,f_{si}\}$对应不确定子集中每个汉字的第 i 位 N 联字频率，这些由事先统计好的 N 联字库中查到。因此，只要计算出上述矩阵，根据 G 值的大小，就能确定出待识汉字 x，即如果存在 r (1<r<s)，使一切 t (1<t<s 且 t≠r)，有 $G_r > G_t$，则 G_r 所对应的汉字为确定汉字。

由此推知：当 n=2，3 时有二联字后处理法和三联字后处理法，其综合评判函数分别为:

$$G = G\ (R,\ f_1,\ f_2\ \ W) = G\ (R,\ W,\ f_{xb},\ f_{ax})$$

$$G = G\ (R,\ f_1,\ f_2,\ f_3,\ W) = G\ (R,\ W,\ f_{xbb},\ f_{axb},\ f_{aax})$$

特别地，有混合型联字后处理法，如混合型 2-3 联字后处理法，这是一种 2 联字和 3 联字混合使用的后处理法。这种方法需要建立的联字库是不确定字在第一位，第二位的二联字库和不确定字在中间的三联字库。其综合评判函数为:

$$G = G\ (R,\ W,\ f_{xb},\ f_{ax},\ f_{axb})$$

显然这种方法比纯三联字后处理法少统计两个汉字，从而减少了很多统计量，同时，它比二联字后处理法多了一项考虑因子 f_{axb}，提高了汉字的确识度，因此，混合型 2-3 联字后处理法具有比纯 2 联字，纯 3 联字后处理法更优越的特点，不失为一种较好的后处理法。如:“办公自动化”中，现假设（化，物）为不确定子集，“自动、动化、动物”都是“动”的二联字，按二联字后处理，“自动化”有可能识别为“自动物”；但按混合型 2-3 联字后处理法一定是“自动化”了。

三、N 联字后处理法的具体实施

3.1 不确定汉字集 X 的生成

将初级汉字识别后的不确定的汉字子集统计起来,生成一个全体不确定字字集 X 及其相应的字频集 W。即 $X = \sum_{i=1}^{n} X_i$ 其中 X_i (i=1,2,3,…,n)为不确定汉字子集，n 为不确定子集数。且 $X_i = (x_{i1},x_{i2},\cdots,x_{imi})$，$m_i$ 为不确定汉字子集 X_i 的汉字数。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一般情况下，各个子集中的不确定字不全是没有重复的，即存在:

$$X_i \bigcap X_j \neq \text{空}(i \neq j\text{且}i,\ j = 1,2,3,\cdots,n)$$

因此，统计过程中，应去掉重复出现的不确定字．现设不确定汉字子集 X_i 事先存在一个队列 Q1 中；统计后的全体不确定字集存在 QX 中；当前正在统计的不确定子集放在 Q 中；则统计生成全体不确定字的算法过程可描述如下:

* 初始化(i = 1; QX = 空; Q = 空;)

* WHILE (Q1 不空) 做{

从队列 Q1 中取一个子集 X 到 Q 中;

i = i+1;

将 Q 中每个没在 QX 中出现过的不确定字按汉字内码大小插入队列 QX 中;

令 Q 为空; }

* 再根据已有资料在 QX 中每个汉字相应的位置上写上字频 W;

3.2 不确定汉字的 N 联字库的设计和生成

为了检索方便，也为了数据的压缩，N 联字库的设计如图二所示:

1、不确定汉字索引库(UnDetermined Index Chinese Library)．此库由若干项组成，每个项存放一个不确定字的索引信息，项序号和内码变化方向一致，即序号小的存放内码小的汉字，整个结构描述如下:

UDIC.LIB = {UDIC(i)|i = 1,2,3,…,不确定字数}

Length (UDIC(i)) = 2 * (N+2); 每个项的长度 (字节单位)

$$UDIC(i) = \begin{cases} \text{汉字内码}hz_i; \\ \text{汉字字频}W_i; \\ N\text{个指针；分别指向该汉字在}N\text{个联字库中的联字首块号} \end{cases}$$

2、各位 N 联字库(N1.lib,N2.lib,N3.lib…NN.lib)．共有 N 个 N 联字库，每个库由若干个定长的 N 联字块 NCB 组成，每个 NCB 由若干项 LZ 组成，每项有 N−1 个汉字和一个联字频率组成，每个块尾附带一个链接指针(Next)，因为联字块是定长的，所以当一个块容纳不下同一个不确定字的 N 联字时，就用下一个联字块 NCB 表示，指针 Next 指示这个后继块号．综上，N 联字库可表示如下:

N.LIB = {N1.lib,N2.lib,…,NN.lib}

Ni.lib = {NCB(j)|j = 1,2,3,…,块数}

NCB(j) = {LZ(i)|i = 1,2,…项数}+指针 Next

LZ(i) = {h(s)|s = 1,2,…,N−1)}+该 N 联字频率 f_i

Length(LZ(i)) = 2 * N;

$$Next = \begin{cases} 0;\ \text{本块}N\text{联字结束} \\ \text{其它正整数；}NCB\text{块号} \end{cases}$$

f_i = 数字;

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

不确定字库　第 1 位 N 联字库　　第 2 位 N 联字库　…　第 N 位 N 联字库

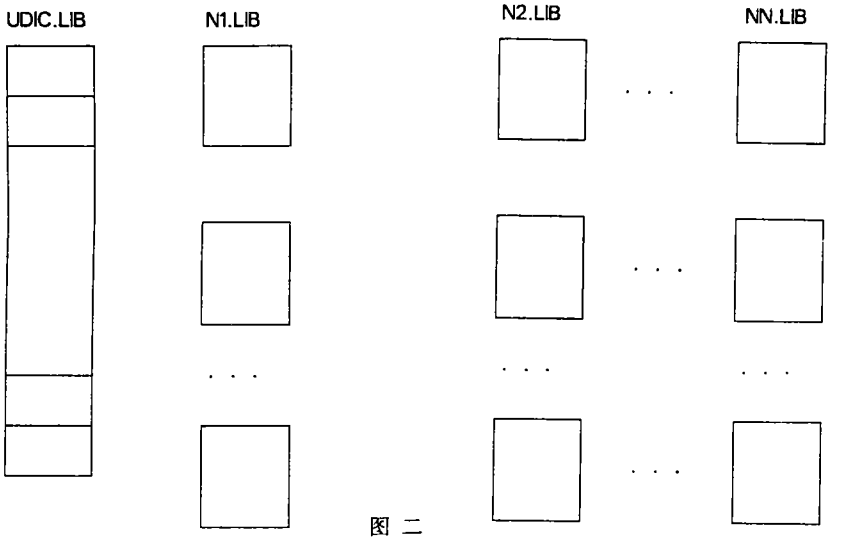


图 二

3、N 联字库的生成。

统计不确定字的 N 联字库时，选定大量具有针对性，代表性，完备性的语言材料，对全体不确定汉字集中的每个不确定字 a_i 找出统计该不确定字各位 N 联字及其出现频率 f_j (a_i (j＝1,2,…,n)，放在各位 N 联字链表中，最后将每个不确定的各位 N 联字及其频率以上面指定的格式存入到文件中。由此，生成了一个不确定字的各位 N 联字及其频率的数据库。在这个数据库中，检索某个汉字的各位 N 联字的过程如下:

第一步: 在索引文件 UDIC 中找到该字所在的项;

第二步: 在该项中取出各位 N 联字的起始块号;

第三步: 在每个相应的联字库中取出相应块号所对应块及其所有后继块（如果存在）的 N 联字及其频率。

同样，这种结构下的不确定字及其 N 联字的增加也相当容易。

3.3 N 联字后处理法的核心程序

定义 3.1 设 $A=a_0W_{x1}(a_1)W_{x2}(a_2)\cdots W_{xn}(a_n)a_{n+1}$ 是初次汉字识别后的结果汉字串，其中 $W_{x1}(a_1)W_{x2}(a_2)\cdots W_{xn}(a_n)$为不确定语段，n 为语段中不确定字数，$a_0$,$a_{n+1}$ 为不确定语段前、后联字，且 a_0，a_{n+1} 属于{确定汉字,空}，而 $W_{xi}(a_i)=\{x_{i1}x_{i2}\cdots x_{imi}\}$是一个 a_i 的不确定字集，m 为 a_i 的不确定字数。

一般说来，不确定字集中都是些形相近的字，由初级汉字识别给出。如（土士），（己已），（人入）等等。因此，在含有不确定语段的汉字串 A＝“是（人，八，入）(类粪)文明”中，$n=2$，$m_1=3$,$m_2=2$,a_0＝是，a_3＝文。

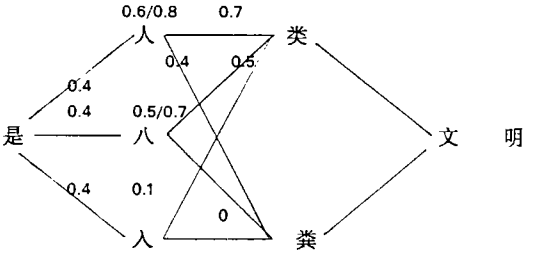
定义 2: 每个不确定汉字的评价函数为各个因素的加权和，即:

$$G_{ij}=K_oR_{ij}+\sum_{s=1}^{N}K_sf_{ij\,s}+K'_0W_{ij}$$

确识规则: 评价函数值 G 最大的那个汉字，被定为确识后的汉字。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

例如：对于语段 A＝“是（人八人）（类粪）文明”，采用 N 为 2 的二联字后处理法，各个二联字频率如图三所示，综合评价函数为 $G_{ij}=R_{ij}+f_{ij1}+f_{ij2}+W_{ij}$ 加权系数都为 1。显然 a_0＝是，a_3＝文，n＝2



图三

因此，确定 a_1 时有：

$$G11=R11+f111+f112+W11=0.6+(0.7/0.4)+0.4+0.8=2.5$$

$$G12=R12+f121+f122+W12=0.5+(0.4/0)+0.4+0.7=2.0$$

$$G13=R13+f131+f132+W13=0.5+(0.1/0)+0.4+0.7=1.7$$

显然 G11 最大，所以 a_1＝人；

同理 G21 最大，所以 a_2＝类。

根据上述定义，我们设初识后的不确定语段 $A=a_0W_{x1}(a_1)W_{x2}(a_2)\cdots W_{xn}(a_n)a_{n+1}$ 存在队列 Q2 中，确定后的语段存入队列 Q0 中，当前正在识别的不确定汉字子集存入队列 Q1 中，则后处理识别算法可描述为：

```
* 令 Q0＝空; Q1＝空; i＝0;
* WHILE (Q2≠空) 做 {
      从不确定语段队列 Q2 中取不确定子集 Wxi(ai)放入 Q1 中;
      i＝i+1;
      IF (Q1 中汉字个数≠1 个) {
          清缺省标志 F;
          WHILE (Q1 不空) {
              从 Q1 中取一个汉字 xij;
              IF(该汉字存在于全体不确定字集 X 中) {计算相应的 Gij 值;}
              ELSE {
                  令 Q1＝空;
                  置缺省标志 F; }
             }
          如果 F 为不缺省且只存在一个最大的值 G 则{
                  将 Q1 中 G 值最大的那个汉字送入 Q0;}
          否则 { 拒识该汉字并将拒识标志送入 Q0 中;}
          }
      ELSE {
```


© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

将 Q1 中汉字送入 Q0;
令 Q1 = 空;}
}

* 将 Q0 中汉字逐个输出;

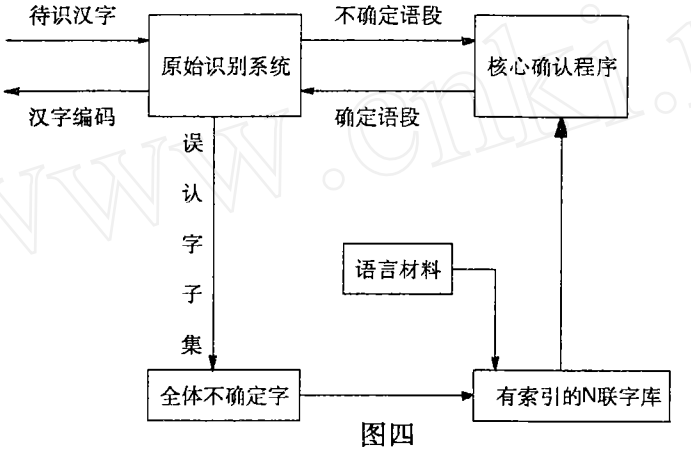
算法中计算 Gij 的过程如下:

* 在不确定字索引文件中取出该不确定字的各位 N 联字块首置 B_i;
* 在各个联字块中取出该不确定字的所有 N 联字 (包括后继联字块中的 N 联字);
* 计算所有 N 联字的 G 值;
* 将最大的 G 值返回;

理论上可以证明该算法是正确的。

四、汉字识别后处理系统

汉字识别后处理系统由后处理核心确认程序、全体不确定字及联字库组成，将这个系统耦合到原先的汉字识别系统[4]上（如图四所示）而成为扩展汉字识别系统。



图四

该系统理论上可提高的识别率:

设初识系统字符集个数为 W，正识别率为 L1；拒识率为 L2，误识率为 L3，误识字库字数 W1，联字条数 N，则后处理中提高的识别率 L4 为:

$$L4 = K * L2 * L1 / (L1+L3)$$

其中 K 是与后处理系统性能有关的系数，如不确定字数 W1，联字条数 N 及后处理算法等等这些因素有关，K 的取值范围为 0−1 之间。

例如：假设 L1 = 90%，L2 = 7%，L3 = 3%，系统性能系统 K = 0.9,则提高的识别率为:

$$L4 = 0.9 * 0.07 * 0.9 / (0.9+0.03) = 6.1\%$$

由此可知：提高确识率的措施有:

要有一部相对全的误认字及其联字字库;

经过单个字识别后的识别率应在一定值之上;

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

初级识别中，在保证正确识别率的同时，应尽量增加拒识率，减少误识率[5]。

五、实验和结论

本文从最初步的N联字的上下文关联利用思想入手，具体从理论上讨论了一种汉字识别后处理方法，形式给出了确认汉字不确定语段的理论算法。在此理论基础上，我们用N为2的2联字汉字识别后处理法做了一系列测试实验，良好的实验结果证明了该方法确能明显提高汉字识别率，因此基于N联字的汉字识别后处理系统，不但实现了理论的设想，而且也证实了结论。

参考文献

[1] 崔国伟、舒文豪、李仲荣，"关于联想式汉字识别后处理方法的研究"，《模式识别与人工智能》，第2卷，第1期，89年3月。

[3] 张忻中、沈兰生，"印刷体汉字识别技术在我国的发展和应用"，《中文信息学报》，VOL. 6，NO.1。

[3] [英] 厄尔曼著，《文字图形识别技术》，人民邮电出版社。

[4] 周昌乐、马希文，"基于互动计算的汉字楷书识别"，《自动化学报》，1992.7,第4期。

[5] 叶乃奉、张忻中、夏莹编，《汉字微型计算机与汉字识别》，机械工业出版社。

A Study Postprocessing Approach to Chinese Character Recognition Based on N-United-Word

Miao Lan Fang　Zhang Sen　Zhou Chang Le

(The computer department of Hang zhou)

Abstract

In order to increase the recognition rate of chinese characters, a postprocessing method to the recognition of chinese characters based on contextual relation of N-united-word has been proposed in this paper. That is the method with which we can further determined the chinese characters that is not determined or refuese to recognize in the priliminary recognized period. First, the main idea and theoretical foundation of this method has been expounded. Second, the structure of the database of the postprocessing based on 2-united-word has been discussed and the theoretical algorithm of the postprocessing of chinese chatacters recognition has been given out. Then, the recognized rate increased in theory has been analyzed. Finally, a system model of N-united-word postprocessing has been given out.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net