试谈利用电子计算机自动编制中文著作字索引(续完)

黄俊杰　陈恩泉　张　普

声码部分基本上依照汉语拼音方案。但是，为了能与西文电传机、五单位穿孔机通用，对键盘上没有的汉语拼音符号作了一些修改。同时，要保证一个汉字的信息不超过计算机的一个字长(709机是48位二进制)，每个汉字的编码则不能多于八个符六字节，因而对个别多于八个字节的音节(包括形码)作了压缩。例如：

(1) 用1、2、3、4代表四个调号；

(2) 当声母L、N与韵母ü结合时，以yu代ü；

(3) 以-m代鼻音韵尾-ng；

(4) 当声母zh，ch，sh与韵母uai，uan，uam结合时，该音节不标调。

由于我们的工作重点主要放在利用电子计算机编制中文著作字索引的程序设计上，因而对汉字的编码及与之相关的整个汉字信息处理系统下的功夫很少，现用的编码系统还存在一些明显的缺点，如：字码太长，归部原则不精，要求译码员掌握标准普通话音系等等。

三、程序系统

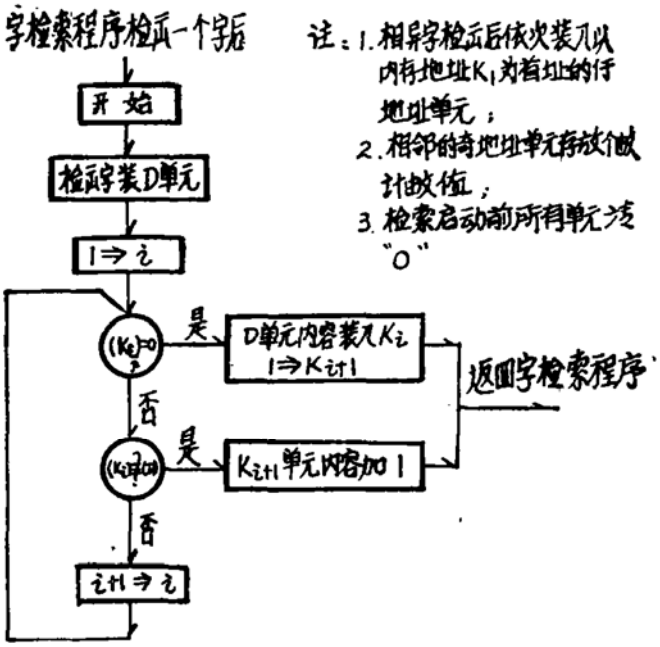
考虑到目前分布面较广的国产709机外部设备中外存(鼓、带)较少，稳定性能较差，且内存容量只有32k，主机为串行方式等因素，为了解决大信息量与记存容量较少的矛盾和检索时间与机器稳定工作时间的矛盾，我们决定采用手编程序。全部按部首、音序、频度检索输出的三个项目的程序约占0.52k。程序设计中有如下问题值得注意：

1. 在目前情况下是不宜采用"字库"存贮方式的。一般认为通用汉字有八千个左右，如果每个汉字占据一个内存单元，另外还必须各分配一个记数单元，单是这一次就占去内存的一半，即使这样，八千汉字的字库也还不能包括一切，必然会有相当数量的库外检字引起的程序中断，从而影响检索的速度。所以我们采用检索过程中见异字而辟单元的方法，实践证明这样不仅可以节约内存单元，而且提高了检索速度，其程序如下图：(见下页)

2. 我们采用的是《拉丁化汉字编码方案》，由二十六个字母及1～4四个阿拉伯数字组成，加上标点符号在内总共约有四十五个不同的符号，以六位二进制数为一个字节足以表征一个字码。我们按709机字符编码规则，以符六为信息的基本输入形式。该编码

方式英语二十六个字母从A到Z刚好是按音序从小到大顺次排列的，这对于我们按音序排列的字频统计程序的简化是极为有利的。

3. 按现在采用的编码系统，字长是不固定的，最长的码字有八个符六字节，刚好是709机一个全字长，但大多数码字均小于八个符六字节，如果严格按一码一存贮单元方



式寄存，加上标点符号，这将浪费太多的内存单元，而且这种信息组织方式另一个最严重的缺点是给信息的输入组织工作带来极大的麻烦。因为按709机的符六输入方式，纸带上没有分单元写入的标识符，机器接收满八个符六字节就写入内存并地址自动加1，不够全字长的码字势必要求补填空格，例：

想 x i a m 3 m x g	X I A M 3 M X G
怕 p a 4 b s x	P A 4 B S X ␣ ␣
,	, ␣ ␣ ␣ ␣ ␣ ␣ ␣

(其中␣表空格符)

同时，如果补空格时稍有差错，将导致以后的信息全部错址。这就给检索信息的纸带穿孔及检校工作提出了极为严格的要求，这对于大部头书籍(几十万字到上百万字)的信息组织工作来说几乎是不可能做到的事情。因而我们采用不定长随意输入方式：码字之间用一个空格符隔开，如码字之间有标点符号相联则此空格符还可略去。这样一来，不但可以大幅度节省内存单元，也使输入信息的组织工作大为简化，而且如果发生码错和穿孔错误等都只影响一个码字，决不会导致成群连锁反应。

由于按部首排列逐字索引"页——行——个数"这种输出格式的需要，我们引进几个特殊的标识符：

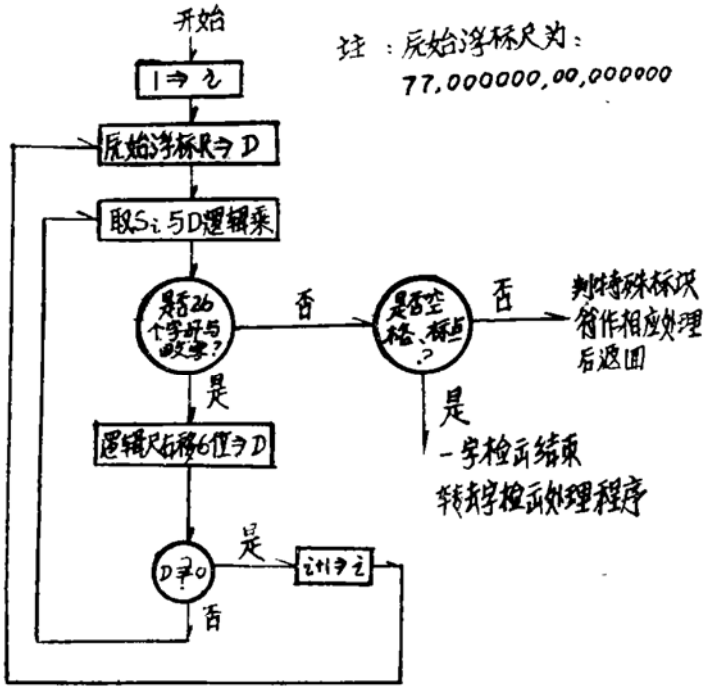
① 跳行码(Ls Code)，在信息的每一行结束时加穿一个跳行码。

② 页终码(Eop Code)，在信息每页最后一字后面穿一个页终码。

③ 分段码(LD. Code)在讯息每一自然段后加一分段码。

④ 块段代码(BLock Code)，对于大部头书籍，我们按 2 万信息单元为一单位分块写入磁带，为了产生块结束讯号，块段结束时写入一个块段代码。

4. 上述信息的结构方式就决定了程序中的检索方式：浮动游标逐字检索方式。其程序如下图：



5. 我们从初步实验中体会文字检索对于计算机有以下几点要求：

① 基本操作要迅速(这是由计算机硬件决定的)。

② 磁记忆的容量要大，存取速度要快。

③ 软件设计要简化。我们之所以采用手编而舍弃应用 709 机所配套的语言程序，除节省内存单元来存放信息这个考虑之外，主要是想获得较快的速度。有些子程序如 2⇒10 子程序，求百分比子程序等在检索过程中反复应用的次数极多，一个单重的时间稍许节省就可较大幅度的减缩检索时间，这是一个十分重要的问题。由于国产 709 机目前连续稳定工作时间还不太长，我们必须尽量在检索速度上予以充分注意。如：2⇒10 子程序在字频统计这个输出项目上，几乎每一个字都需应用一次，如果它的速度提高 T 抄，检索的常用汉字为四千个的话，那么节省的总时间就是 4000×T 秒，这是一个很可观的数字。

④ 访问外存的次数要尽量减少。原因之一是外存与主机交换信息的时间较长，次数太多影响检索速度。另外一个原因是，在目前我国外部设备的技术状况下，交换次数愈多，出错的机会也就相应增加。在我们试验性工作中以按部首排列的逐字索引输出较为难办，因为 709 机的内、外存较少，我们只得一个一个部首来进行，这样记录在带上

的信息就得读取一百九十次(部首个数为一百九十)这样频繁的调带工作不但误事而且误时。如我们的试验表明，按频度输出，按音序输出这二个项目所用的全部时间还不到按部首检索输出时间的三十分之一，这是一个急待解决的问题，我们准备拟订新的方案再进行试验。

利用电子计算机进行语言研究，有许许多多工作要做。我们所做的仅仅是一点实验性工作，即使这样，也还存在着一些问题需要我们去解决，如编码系统，程序设计都有待进一步完善。今后，我们将在已取得的研究成果的基础上，扩大研究范围，拟订出新方案再进行试验。

(上接第80页)

因为蝙蝠的形态、习性虽有很多方面和鸟类一样，然而鸟类是卵生的，而蝙蝠不具有卵生这种属性。这也是以对象内部的机制作为鉴别标记，所以结论也就较为可靠了。

第二，对比所发现的差异属性(d)，如果正是某类对象的独特性（相当于定义中所讲的"种差"），那么结论就非常可靠了。例如每种化学元素都有其独特的光谱，所以，称它为元素的标记光谱。凭标记光谱就能非常准确地辩别元素。而且，只要分析一下各个矿石样品(化学物)的光谱，也就可以把极为相似的不同矿石区别开。不过，真正把握一类对象内在的独特属性，这是需要经过长期艰苦的努力才能做到的。

应用对比鉴别推理的关键是要找出作为对象标记的属性。如电子和正电子无论质量的绝对值或电量的绝对值都是相同的，也都是稳定的粒子，只是电荷相反。当它们进入威尔逊云雾室时，在强磁场的作用下，就会留下弯曲的径迹。对比两者留下的径迹，弯曲的方向恰好相反。由于找出此种径迹作为标记，这样，也就可以对电子和正电子作出鉴别推理了。从一九三二年发现正电子后到现在，物理学家又先后发现了一系列的"反粒子"。正象普通的正粒子组成普通的物体一样，"反粒子"也能结合成"反物体"(亦称"反物质")。如普通的氢原子是由一个质子和一个电子组成的，而"反氢"原子是由一个反质子和一个正电子组成的。那么，在无限宇宙中，有没有由"反粒子"组成的"反星体"呢？如果有的话，到目前为止，人们还无法把它与普通的星体加以鉴别，因为还找不出可供辩别的标记属性。离我们遥远的"反星体"(如果存在的话)，它们的引力效应和它们所产生的光与普通星体是完全一样的。因而，直到现在还无法作出鉴别的推理。找出普通星体与"反星体"的鉴别标记，这正是科学工作者所努力探索的事。如果能够确定已知的普通星体都具有属性 a、b、c、d，而且又观测到遥远的某些星体只具有属性 a、b、c，不具有属性 d，那么这就非常有益于"反星体"的发现工作。

以上是我们关于归纳划类推理与对比鉴别推理的一些不成熟的观点，难免有欠妥之处，请读者批评、指正。我们觉得，重要的是应当不断地从认识史、科学史的实际进程中作出逻辑的概括和总结，积极地开展科学逻辑与科学方法论的研究。至于这种研究工作中会出现这样的或那样的见解，出现这样的或那样的误差，那是一点也不奇怪的。