

●黄水清 侯汉清

中国图书馆学报(季刊)1991年第3期
ZHONGGUO TUSHUGUANXUEBAO

汉语 PRECIS 在计算机上的实现 ——汉语 PRECIS 款目生成系统研制报告

PRECIS(Preserved Context Indexing System——保留上下文索引系统)是60年代末至70年代初产生于英国的一种新型字顺主题索引系统。它是一种人工辅助机编索引系统，在理论上适用于任何自然语言。PRECIS 目前已获得世界性的声誉，被誉为“可能是自 1876 年克特首次拟定主题标引规则以来最重要的建树。它从一个全新的角度对图书及其他文献进行主题标引”⁽¹⁾。除了英语，它还成功地应用于德语、法语、丹麦语等自然语言。从 1987 年起，我们开始对汉语 PRECIS 进行研究，并设计了完整的汉语 PRECIS 职能符号系统。在此基础上，实现了汉语 PRECIS 的计算机自动生成款目过程。

一、汉语 PRECIS 及其特点

汉语 PRECIS 是在英语 PRECIS 的基础上，遵循汉语句法规则，将英语 PRECIS 中不适用于汉语的职能符号及其使用规则加以调整、修正而得到的专门用于汉语文献的主题索引系统⁽²⁾。它的工作方式为：由标引人员写出代表文献主题的主题描述句。并用职能符号标记主题描述句中各主题词在句子中所承担的逻辑功能，得到代表文献主题的汉语 PRECIS 词串。将词串输入计算机，计算机将自动完成款目的生成、排序、书本式索引的打印等工作，并可形成机读数据库，以备机检之用。

汉语 PRECIS 的款目为双行形式。款目中的各个词都处于既定的上下文中，打破了传统的主、副标题型款目的限制，可以容纳众多的主题词。这样，标引的专指度便不受制于款目形式方面的缺陷。而完全取决于文献主题本身。如果必要，词串中的任一主题词都能轮排到款目的首词位置作为检索入口点。而不妨碍生成款目的准确性。

索引款目的可读性强。这首先表现在汉语 PRECIS 的引用次序与自然语言贴近，款目中的上下文对主题词的限定与自然语言中语境对语词的限定完全一样。此外，汉语 PRECIS 的轮排过程不影响款目意义。轮排生成的款目与自然语言贴近。标引过程中使用的职能符号在款目中一个也不出现。按照规定的判读顺序阅读轮排款目，就可得到近似自然语言的语句，且不改变主题意义。

汉语 PRECIS 标引能力强。可以从标引深度和标引一致性两方面考察标引能力。汉语 PRECIS 的平均标引深度大于目前通行的中文主题标引方式。在标引一致性方面，由于汉语 PRECIS 职能符号系统严谨完备，规则详细统一，能标引各种类型的主题，减少了标引员的主观臆想，保证了标引的一致性。汉语 PRECIS 的良好标引能力特别表现在对复杂主题的描述和处理上。这是其它索引系统难于企及的。

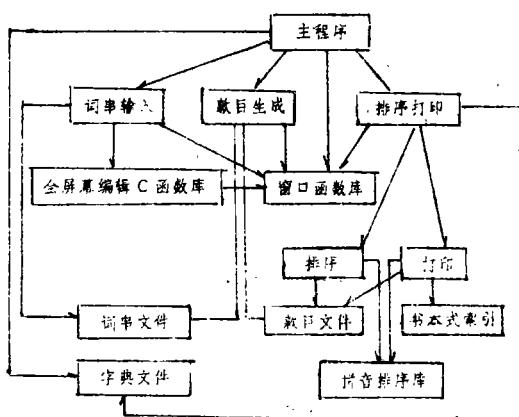
最重要的是，汉语 PRECIS 是一种人工辅助机编索引系统。它由人工完成需要智能的主题分析和标引工作，而由计算机完成

轮排款目的生成、排序、书本式索引的编辑打印等纯事务性的工作。既充分发挥了人、机各自的优势，加快了编制速度，又保证了索引质量。

二、系统功能与系统结构

汉语 PRECIS 软件系统以 C 语言为开发工具, 采用贝尔实验室的 Lattice C V3.0 编译程序在 Sun286 微机上开发成功。在支持 CC DOS 的 8088、80286、80386 各类 IBM-PC / XT 或 AT 兼容机上均可运行, 稍作修改便可移植到 XENIX 运行环境。

从功能上看，系统由词串输入、款目生成、排序打印三个子系统组成。其中的词串输入子系统采用全屏幕编辑工作方式。为此，开发了一通用全屏幕编辑 C 函数库作为支持软件。在词串输入过程中，为检验输入词串的正确与否，需将由词串生成的各条款目显示在屏幕上，以供输入人员校验。词串输入子系统与款目生成子系统之间存在调用与被调用的关系。系统结构图如下：



启动系统后，主程序的运行结果是从字典文件中读入预先设置的系统参数，并在屏幕上生成如下形式的主菜单：

- 2 款目生成
 - 3 排序打印
 - 4 退出

然后，根据用户键入的选择，依次调用各个子系统。子系统调用完毕，将返回主程序，重新显示上述主菜单，直至选择 4 结束系统运行。系统的整个流程与目前大多数检索刊物的字顺主题索引编制过程是一致的。最后得到的是按拼音排序的书本式汉语 PRECIS 索引。

三、通用全屏幕编辑 C 函数库

原始数据的录入过程可以有行编辑和全屏幕编辑两种方式。其中当然以全屏幕编辑为佳。但由于全屏幕编辑程序的开发涉及键盘和显示器控制，而大多数微机上的高级语言编译程序并不支持这种功能，故在许多软件系统中程序员不得不转而采用行编辑方式。Lattice C 的标准库函数中，也无屏幕处理函数。在这里，我们为了能在全屏幕编辑方式下进行词串的输入、修改等操作，参照 XENIX 下的 C 语言窗口函数库，采用 CC DOS 软中断调用方式，专门开发了与之兼容的窗口函数库。这个窗口函数库，具有开窗口、删除窗口、刷新窗口内容，清除窗口（或整个屏幕）、移动光标至指定位置、获取光标当前位置坐标、插入字符（或字符串）、删除字符（或字符串）、读字符（或字符串）、写字符（或字符串）等功能。其调用方式亦与 XENIX 完全一致。

我们的全屏幕编辑 C 函数库即在此窗口函数库的基础上开发。它可以根据从字典文件读入的一幅窗口描述。在此窗口内，上下左右自由地移动光标到指定的输入位置（窗口内的提示信息显示位，光标将自动跳过），对输入数据作添加、删除、插入、更新、读取等操作。该函数库还支持对一个以上的窗口同时进行操作。

窗口函数库的实现是采用软中断调用方式，它只依赖于 CC DOS，与机器硬件性能无关。不论是 CGA、EGA 还是 VGA 显示卡，不论是低分辨率、中分辨率还是高分辨率屏幕，不论是 10 行、16 行还是 25 行的 CC DOS，该函数库都能正确运行。相应地，在它之上开发的全屏幕编辑 C 函数库也是一通用函数库。C 语言规定，编译系统的库函数与用户的自定义函数性能完全一样。这两个 C 函数库可以看作是对 Lattice C 编译程序的扩充。

在本系统中，实际屏幕的第一行被置为公用提示行，用于显示各级子菜单和程序运行过程中的其它提示信息。并接收人机对话过程中键入的选择。若当前屏幕正处于全屏幕编辑状态，则实际屏幕的倒数第 2 行被置为编辑键提示行，提示各编辑键的功能。公用提示行、编辑键提示行采用反白显示方式。窗口内的提示信息以高辉度显示，用户键入数据以正常方式显示，以示区别。

四、词串输入子系统

词串输入子系统具有词串的添加、更新和竖式格式的打印功能。它在全屏幕编辑 C 函数库的基础上开发成功。

在主菜单下选择进入词串输入子系统后，系统首先在公用提示行显示“词串文件名：”并等待输入。若词串文件为新文件，系统还将在公用提示行显示“最小词串号：”，读入的最小词串号将作为词串文件中第 1 条词串的词串号写入词串文件。

词串编辑过程中，每幅屏幕处理汉语 PRECIS 词串中的一个复合词。其屏幕格式如下：

PgUp 翻页至上一复合词 PgDn 翻页至下一复合词 F1 插入一复合词 F10 结束全屏幕编辑

当前词串号为：

〔论题连件〕〔中心词职能符号〕〔是否领词〕〔是否替代〕〔替代数〕
 〔中心词值〕
 〔区分词 1 \$〕
 〔区分词 2 \$〕
 〔区分词 3 \$〕
 〔区分词 4 \$〕
 〔区分词 5 \$〕
 〔区分词 6 \$〕
 〔区分词 7 \$〕
 〔区分词 8 \$〕
 〔区分词 9 \$〕
 〔下读连件 \$ V〕
 〔上读连件 \$ W〕

↑ 上移 ↓ 下移 CR 下移 ← 左移 → 右移 BS 删左侧字符 Del 删当前字符 Ins 插入

对复合词作编辑时，若按 F1 键，则系统认为用户需要在当前复合词的之前或之后插入一个新的复合词。此时，公用提示行显示如下形式的子菜单：

(0) 插在当前复合词之前 (1) 插在当前复合词之后 请选择：

选择 0 或 1，屏幕上显示一幅新的复合词输入窗口，各数据输入段都被清空，以备用户键入新数据。

PgUp 和 PgDn 键分别用于将当前复合词之前或之后的复合词调到当前屏幕，成为新的当前复合词。若当前复合词已是词串中第 1 个复合词，则按 PgUp 键不改变当前屏幕。同样，若当前复合词已是词串中最末一个复合词，则按 PgDn 键也不改变当前屏幕。利用 PgUp 和 PgDn 键，用户可以方便地对当前词串中的任一复合词进行操作。

在“〔论题连件〕”的输入位置，按汉语 PRECIS 规则，应该输入 X、Y 或 Z。若在此位置输入空格，则表示将当前复合词从词

串中删去。

若要结束词串编辑，只须按 F10 键，系统则将刚编辑完毕的当前词串呈竖行形式显示在屏幕上，并调用款目生成子系统生成对应的款目。这些款目可逐条在屏幕上显示。这样做的目的是为了用户能最后检查屏幕上的词串和款目是否完全正确。这期间，系统还将检查当前词串各职能符号的合法性；若发现非法职能符号，系统在公用提示行给出警告。

若用户认为屏幕上的词串已无须再作修改，可按 F10 键承认这次词串编辑。否则，可按 F1 键，对当前词串再作编辑。

词串的添加和更新分别通过调用上述过程而实现。在词串添加中，系统将定位在词串文件尾，并将词串文件中最后一条词串的词串号自动加 1，作为新词串号。以后，每输入一条新词串，词串号就加 1，直到输入一空词串，结束本次词串添加过程。在词串更新中，系统首先在公用提示行询问欲更新的词串号，根据键入的词串号从词串文件中读入相应的词串，并将该词串切分成若干复合词，用于初始化复合词输入屏幕，再对其进行编辑。

词串输入子系统还可按竖行形式打印词串文件中的指定词串。执行打印词串功能时，系统在公用提示行依次询问欲打印的最小词串号和最大词串号。位于最小词串号和最大词串号之间（包括它们本身）的词串将呈竖行形式依次打印出来。

五、款目生成子系统

款目生成子系统从词串文件中的第 1 条词串开始，逐条读入词串文件中的词串，并按汉语 PRECIS 的职能符号生成款目，写到指定的款目文件中。

调用款目生成子系统后，公用提示行首先显示“词串文件名：”，并等待输入。如果

系统按读入的词串文件名打开词串文件失败，则在公用提示行显示错误信息，并返回主菜单。如果文件打开成功，则公用提示行再显示“款目文件名：”，并等待输入。如果此时输入的款目文件名指向一已存在的文件，则该文件将被清空。

款目生成过程中，系统将再次检查词串中各职能符号的合法性。若检测到非法职能符号，将在公用提示行显示出错信息，指明是哪类职能符号出错，并返回主菜单。

六、排序打印子系统

该子系统包含两项功能：对款目文件按拼音排序；将排序后的款目文件呈分栏书本索引形式输出。执行该子系统时，系统将首先从当前驱动器读入拼音排序库。在拼音排序库内有国标一级和二级字库所有汉字和符号的拼音排序码。若系统读拼音排序库失败，则显示出错，并返回主菜单。

这里，若将拼音排序库换成其它形式的汉字排序库，如笔画、笔顺，则系统对款目文件按笔画、笔顺排序。

排序程序采用了置换-选择排序生成初始归并段的多路归并技术，内排序采用快速排序算法。

设程序在内存中为排序开辟了能容纳 W 个款目记录的工作区。W 的值是作为系统参数定义在字典文件中的。由算法分析可知，置换选择排序生成的初始归并段平均长度为 $2W$ 。初始归并段生成结束后，这个大小为 W 的工作区又被用于多路归并。平均来说，一趟归并最多可排序的款目记录数为： $2W \cdot W = 2W^2$ 。若在字典文件中置 W 为 256，即工作区可容纳 256 条款目记录（在一般 IBM-PC 兼容机上，内存空间足以开辟这么大的工作区）。此时，一趟归并平均起来可排序的最大款目记录数量为 131072 条。程序执行过程中，生成初始归

并段需读整个文件一次。作一趟归并又需读整个文件一次。13万条记录排序只需读完整个文件两遍。

当归并段的数量大于 W 时，排序程序先对从 1 至 W 的归并段作一趟归并，再对 W+1 至 2W 的归并段作一趟归并。……依此类推。这些过程在读一遍文件过程中完成。每对文件中的归并段作一遍归并，实际上包括了数趟 W 路的归并。设归并前文件中有 n 个归并段，则归并一遍的归并段数量为 $\lceil n / W \rceil$ 。

另外，设经置换—选择排序后得到的初始归并段总数为 N，则可以推算出整个排序算法的时间复杂程度为 O。这里不包括归并过程中的访外时间。访外的次数在上文已作了分析。为了最大限度地减少访外造成的时间延误，在归并过程中采用了举行缓冲区技术，其并行缓冲区个数亦为 W。

排序之后便可输出书本式索引。书本式索引的格式可以由用户自行设定。也可采用系统缺省值。当系统显示“设定打印格式？(Y/N)”时，若用户回答“N”或“n”，则采用下述系统缺省值：

行宽	80 个字符
栏目	2(双栏)
行数	55

打印格式设定后，系统将询问要输入的款目文件名，然后显示“输出到磁盘？(Y/N)”。若用户回答“Y”或“y”，则生成的书本式索引以扩展名“PRT”写到磁盘，否则直接输出到打印机。

打印程序还可以接收一组打印头信息作为书本式索引第 1 页的最初几行输出。这样，可以在书本式索引的开头得到“1989 年度主题索引”之类的标题。系统自动将这些打印头信息归中打印。

七、数据测试与遗留的问题

我们以 1985 年第 6 期《中国农业文摘·土壤肥料分册》(共收 239 篇文献)和 1988 年第 6 期《水稻文摘》(共收 218 篇文献)作为期刊文献测试数据。同时从 1987 年北京图书馆统编卡片中任选了 200 张(从中又剔去小说、剧本等文艺作品书籍 34 张)作为图书测试数据。对以上三份测试数据分别用汉语 PRECIS 方式作标引，得到相应的词串。然后运行系统，依次执行输入、生成、排序、打印等功能。全部词串都准确无误地生成了对应款目，分别输出了三份书本式汉语 PRECIS 索引^[3]。

目前系统还不具备检索功能，也没有配备相应的文献数据库，不能实现汉语 PRECIS 词串的计算机检索。检索及词表管理系统仍在研制中。

参考文献

- (1) Wellisch, H.H., The PRECIS Index System: Principles, Application and Prospects, New York: Wilson, 1977.
- (2) 黄水清。汉语 PRECIS 系统——职能号与标引技术：(学位论文)。北京：北京大学图书馆学情报系，1987
- (3) 黄水清，侯汉清。用汉语 PRECIS 标引农业文献的探讨。农业图书馆，1988，(3)

(作者单位：南京农业大学图书情报系。来稿时间：1990.11。编发者：刘喜申。)

Library Science and with the Bulletin of China Society of Library, the initial issue of which was published in 1957, as its predecessor, has totally issued 21 volumes, 94 issues and published 1800 pieces of articles by 1,199 authors. The average amount of citations has been 1.2 items and articles written by personal co-authors have accounted for 8.7% with 4.6 times going-up and the highest distribution has been of 28,000 copies with over 14,000 in 1991. At present, there is an encouraging change in the distribution of articles of various disciplines issued, the patterns of subject matter, the originality and the depth of research. The Bulletin has also made a due contribution to the library construction of this country in all aspects of academic orientation, the pushing forward both with the build-up of Chinese information retrieval language and the studies of bibliography, book science, bibliotics and textual criticism as well as material cumulation. The Bulletin is one of the library and information journals more widely distributed with a quality increasingly going up and at the same time leaps to the list of the 10 journals of worldwide and vast distribution. 10 references.

China Society of Library Science -- Journals

Periodicals -- Research reports.

G250-55

The Way to Make the Chinese PRECIS Computerized / Huang Shuiqing and Hou Hanqing // Bulletin of the Library Science in China / China Society of the Library Science.-1991,17(3).-29~33

The article is a research report on the preparation of Chinese PRECIS (Preserved Context Indexing system) entry production system. The system is composed of three subsystems, word--serial input, entry production and order--in--array print, among which the word--serial subsystem adopts an edit routine of all--edit screen and uses C function stock as the supporting soft--ware which applies to all kinds of compatible machines of IBM-PC / XT or AT. The system has also given a data--test by means of "Chinese Agriculture Abstract, Soil Fertilizer Fascicle", "Rice Abstract" and the centralized cataloging cards of the National Library of China, and this results in a book--form Chinese PRECIS Indexing with a very good production quality. 1 illus. 3 references.

Computer system -- Preparations

Preserved Context Indexing system -- Chinese Language

Subject indexing system -- Computer-produced indexes

G354.4

A Study of the Documentation Activities and the Pedigree of the well-known "Changshu South Zhang Family" / Zheng Weizhang // Bulletin of the Library Science in China / China Society of the Library Science.- 1991,17(3).-34~40

Zhang Haipeng (1775-1816) and Zhang Jinwu (1787-1829) of Changshu city being the famous book engravers, book collectors, bibliographers and bibliotists in the mid-Qing Dynasty, bore a direct relation to their family background. The Zhang family who began to engage in documentation activities traces back to the time between Zhengde and Jiajing of Ming Dynasty (in the early sixteenth century). Although the whole family often suffered from hunger, Zhang Wenlin (1482-1548), their forefather, spared no expense in engraving and his works were excellent. Engaging in engraving even earlier than "Mao Jinji on Ancient Pavilion