汉字情报检索系统CCIRS研究报告*(I)

姚卿达
(计算机科学系)

汉字情报的计算机处理是情报检索系统研究中的一个重要课题。近年来，国内外许多学者致力于这方面研究，创造了不少成果。目前，汉字输出以及汉字发生器的研究较为成熟，而在汉字输入方面还存在许多需要解决的问题。

我们根据汉字情报检索系统的技术需求，在充分利用已有成果的基础上，研制了一种实用的汉字输入方案，其特点是利用图形装置，进行笔触字键书式的整字输入，并在M—150计算机上初步建立了一个较完整的联机汉字情报检索系统CCIRS。本文是关于CCIRS系统功能与结构的研究报告。

§1. 系 统 功 能

CCIRS的总目标是实现汉字情报的计算机存贮、检索、编辑与处理，建立功能较全的联机汉字情报检索系统。系统采用中华人民共和国国家标准汉字字符集及其编码[1]，利用通用计算机设备，通过软件实现汉字的输入与输出，终端用户和中心用户可以选用笔触字键书方式或操作字符代码键方式，将汉字信息送入计算机；同时可以方便地检索所需要的情报，以极其直观的汉字记录形式输出。

根据系统目标，从系统功能设计开始，采取“自顶向下逐步求精”的设计方法，按功能划分模块，对各模块规模、接口、属性以及编制语言作出规定，然后进入算法设计与编码调试，直到运行为止。

CCIRS的功能可以划分为四个方面（见图1）：

一。情报库作成、追加与维护

情报库是指存贮在磁盘存贮器上的情报记录之集合，通过文件（亦称数据组）和数据库而组织起来。CCIRS情报库混合使用BSAM(Basic Sequential Access Method)、BDAM(Basic Direct Access Method)和PDM(Practical Data Manager)等存取方法，利用虚拟存贮操作系统VOS2所提供的数据管理功能，设计了下列程序：

---

*参加本工作的有本校计算中心校外数据站人员；
日本日立系统开发研究所日本语情报检索系统[2,3]及日本中央电子株式会社应用系统部在微型计算机CEC800上建立的小型汉字系统[9]的人员向作者介绍了他们的工作。

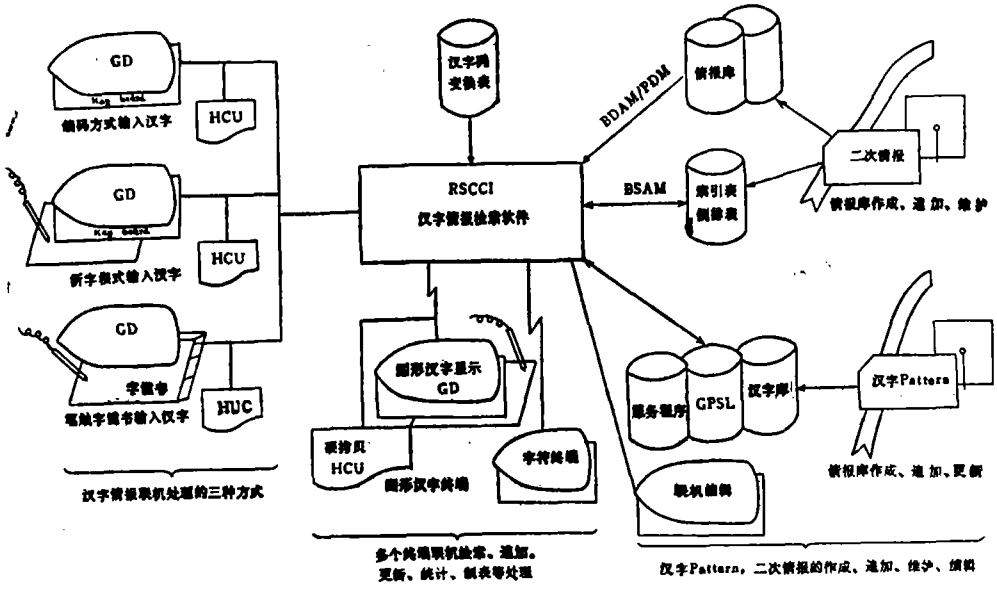


图1 CCIRS功能构成图

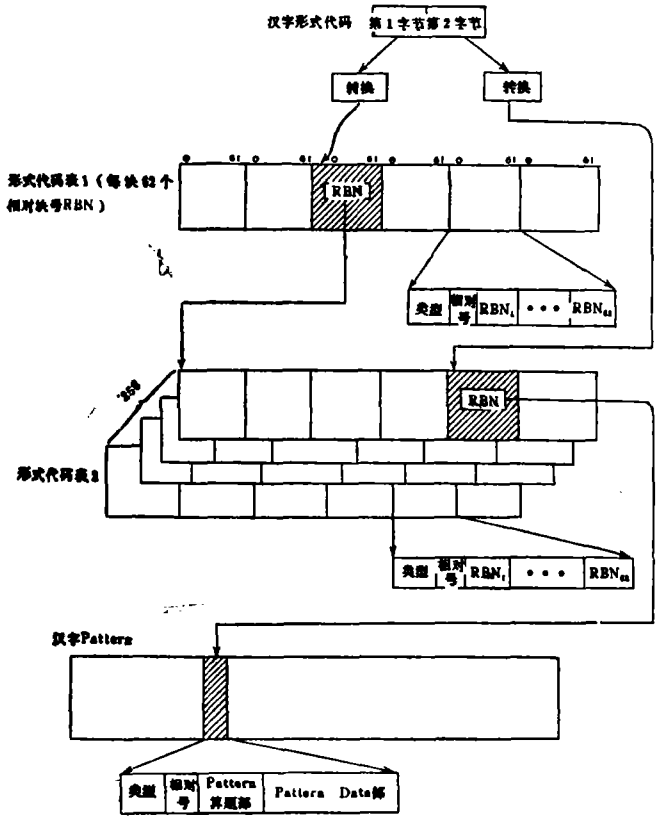


图2 汉字库结构

- BSAM数据组作成、追加与更新
- BDAM数据组作成、追加与更新
- PDM数据库模型DBM作成
- PDM存取说明块ASB作成
- PDM数据库数据作成、追加与更新

为了提高系统效率，这些程序均以批处理方式运行。

二、汉字库作成、追加与编辑

CCIRS汉字库由全体汉字模型数据（Pattern）、汉字形式代码表1和2所组成。仿照GPSL[7]方法，建立了直接编成的汉字库，其结构方式如图2。汉字库按我国标准汉字字符集的字形与编码，制作汉字Pattern，并将

GPSL[7]中约2000个日本汉字数据加以改造，收容到CCIRS库里。有关汉字库的程序：

- 汉字库作成与格式化
- 汉字追加
- 汉字Pattern更新
- 汉字代码转换
- 汉字Pattern联机编辑

CCIRS汉字库备有国家标准的6349个汉字，其中常用的一级汉字3930个，二级汉字2419个。

三、联机情报检索

情报检索是从存贮的大宗情报中及时地找出必要的和充分的情报，并以汉字画面展现在终端屏幕上或由硬拷贝输出给情报要求者。CCIRS试验中曾以计算机科学系学生资料作为情报库内容，进行学生资料管理，用户可以在终端上按学生证号，或按学生名字的汉字代码，找出该生有关材料；也可以用输入笔在字键书上点出一人名字而将其有关材料显示在面前；还可以按某种条件或条件组合式，找出具有某些属性的材料，并以一定的格式列表输出。

处理联机检索的程序有：

1．直接搜索程序
2．倒排表搜索程序
3．数据库检索程序
4．条件式分析与检索程序
5．列表程序
6．追加记录程序
7．更新记录程序
8．删除记录与压缩程序
9．统计运算程序
10．特定动态分析程序
11．联机会话程序
12．错误处理与恢复程序

四、汉字情报的输入、输出

系统提供了汉字编码输入方式与笔触字键书整字输入方式。编码输入对于经过专门训练的终端操作人员是很有效的，而笔触输入对于临时用户或一般业务人员则提供极大方便。设计中还考虑了整字输入与拆字根相结合的方式（图1），将根据实际需要逐步编入系统。编码输入利用字符键盘，采用国家汉字标准编码[1]或电报编码输入汉字信息。国家标准编码分两级，一级汉字按音序排列（同音字按笔画多少排列），二级汉字按部首排列(部首相同者按笔画多少排列)。笔触整字输入是利用图形输入板（Tablet）加上专门设计的字键页作为面罩（Mask），通过软件实现汉字输入。处理汉字输入、输出的程序有：

1．标准编码输入与转换
2．电报码输入与转换
3．字键书页码控制与功能键处理
4．字键书数据键输入与转换
5．汉字画面程序
6．记录与短文编辑
7．字根分析与合并

还有各种代码表。字键书分通用页与专用页两类：

（1）通用页排列常用汉字；

（2）专用页排列某项业务处理中常用词汇与单字，按使用习惯与用字频度排列位置。为了使用方便，允许各页有少量字重现。

§2. 系 统 结 构

CCIRS按模块结构方式组织系统内的程序模块，全体模块分本体模块与支持模块两类。本体模块照前面所述的四部分功能划分，对每一模块的作用、规模、接口、属性和算法作出规定后，分别用COBOL、FORTRAN、ASSEMBLY语言写出，有的部分还使用PDM/DML（数据操作语言）、DBDL（数据库定义语言）和ASDL（存取说明定义语言）描述，经过语言处理程序、连结程序以及有关的实用程序处理后，作成装填模块保存于模块库中，以备调用。

CCIRS在虚拟存贮操作系统VOS2的支持下运行，它能适应于批处理BATCH、分时TSS（Time Sharing System）以及实时联机TMS—3V（Transaction Management System3V）三种处理环境。为了充分发挥VOS2功能，使其同时能照常处理其它作业，提高系统效率与性能，我们根据系统目标、CCIRS特点以及资源情况，设定并生成了一个规模合适的运行系统，运行时要求：主存容量等于1MB，虚存容量等于6 MB、辅存容量不小于280MB。

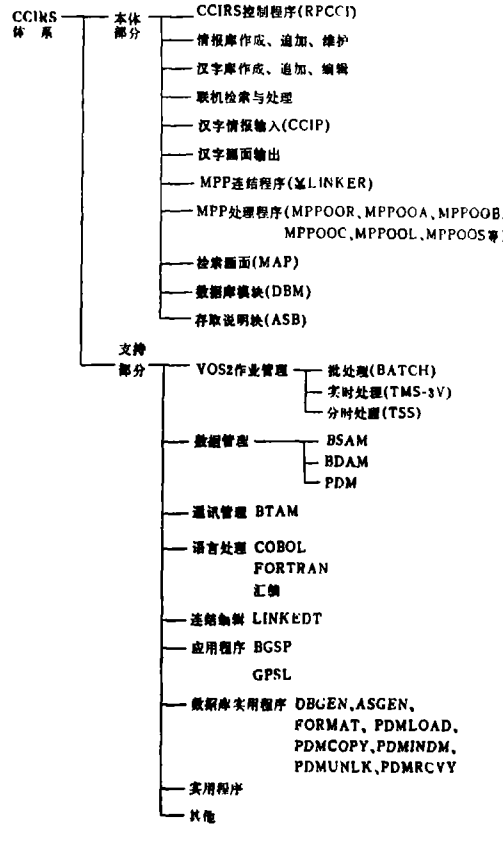
由于CCIRS具有多种功能，又适于不同的运行环境，满足各种情报要求，所以其模块之间接口关系显得较复杂，现将它分述如下。

一、CCIRS 体系(见右表)

二、批处理环境中的结构方式与信息路径

CCIRS 在运行时的信息路径反映了信息的处理流程与工作原理。在批处理环境中，结构方式与信息路径如图 3 所示。图中凡带有细实线的框为本体部分，虚线表示信息路径。检索处理中，信息转换与处理流程如图 4：

处理流程就是开动各程序模块对信息逐步加工，各个程序模块之间的控制转移、入口信息、出口信息、返回代码、信息详细格式与传递规程，均在系统设计时作了规定。对于支持模块来说，则遵从相应的约定与接口。



(支持部分只是将那些与CCIRS有直接关系的列出)

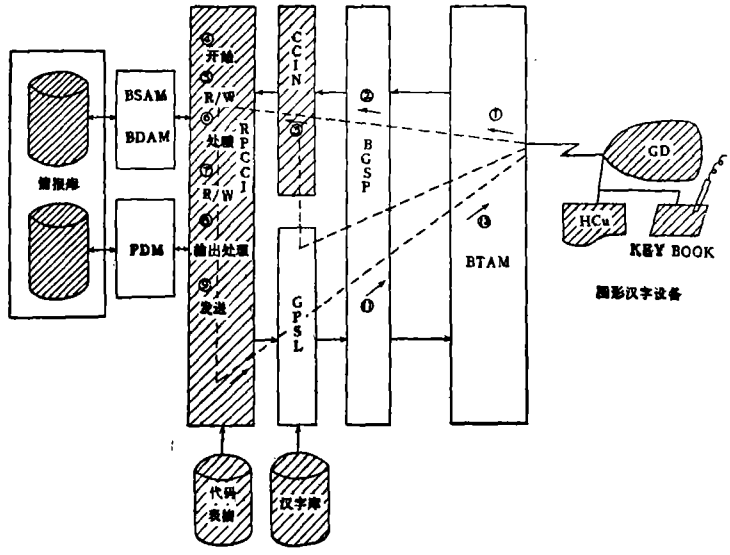


图3　Batch环境中CCIRS结构方式

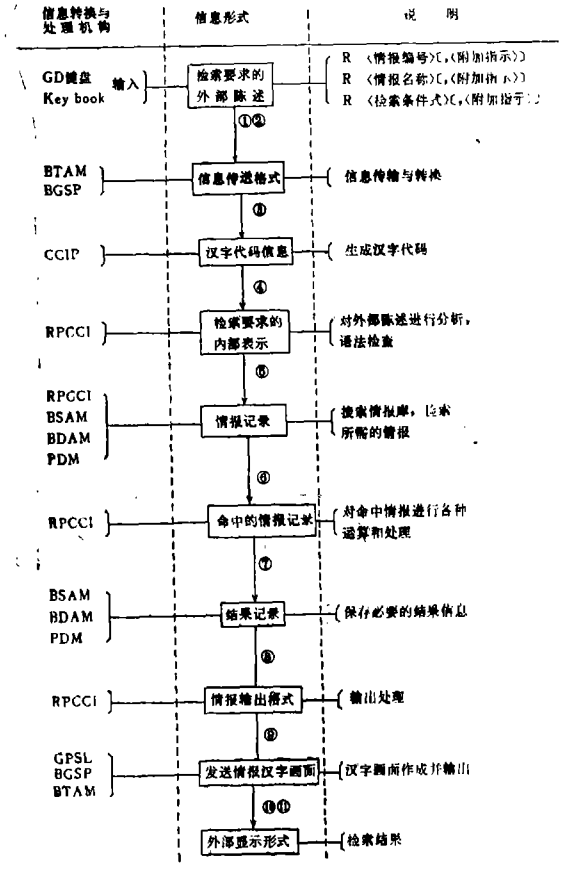


图4　信息转换与处理流程

三、分时与批处理混合形态中的结构与处理方式

CCIRS不仅提供汉字终端接口，而且设计了字符终端接口（结构方式如图5）。对于汉字终端而言，工作方式与前面所述的BATCH方式类似，也可以在TSS支持下，运用TSS机能和RPCCI子命令，进行Q—A检索。对于字符终端，则全在TSS支持下工作，除了使用TSS命令外，增加了RPCCI子命令和Q—A检索功能。当终端接通后，用命令 GPGO 'RPCCI' 便可接通CCIRS，经过设备选择以及必要的初始对话后，系统提示：

Select one of the following options:

1—NAME KEY RETRIEVING　2—NUMBER KEY RETRIEVING
A—ADD RECORD TO FILE　B—BATCH RETRIEVING
C—CHANGE RECORD　D—DELETE RECORDS
E—EDIT TEXT　T—TABLET KEY
Q—QUIT

用户根据需要选其中之一，并按开头的单字命令符回答，使系统转入相应的处理，同时开始进一步对话。

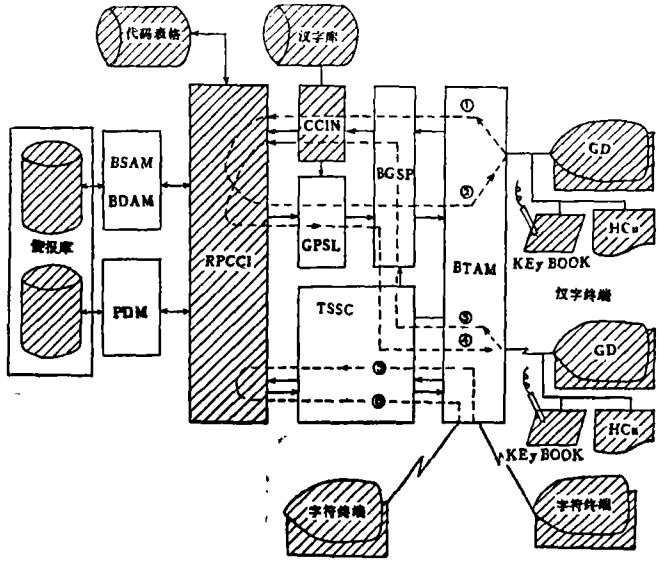


图5 TSS与Batch混合形态中CCIRS结构方式

四、实时联机环境中结构方式与信息路径

实时联机环境中CCIRS结构方式与信息路径如图6、图7所示。字符终端在TMS—3V支持下工作，情报检索以及种种处理由相应的业务处理程序MPP（Message processing program）实现，系统设计中安排了多个MPP接口，包括起始、关闭、检索、追加、更新、统计、列表等的MPP以及成批处理的BMP，以适应多个终端实时联机检索的要求，所有MPP都通过连结程序￥LINKER与TMS—3V连接，￥LINKER

是可扩充的。系统中设计了多种画面，供¥LINKER和MPP调用。连结程序¥LINKLER与MPP具有可再执行与可重入之属性，具有开销少、效率高、对外来事件响应快等特点。由于字符终端无汉字功能，为此设计了检索结果文件，存放必要的结果，委托汉字设备输出，或由汉字终端选择输出。对于汉字终端，其工作原理与批处理环境相同，但数据存取方面作了只读（Readonly）的限制（只允许汉字终端从情报库与结果文件中读取记录），进行检索与列制表格处理，不允许追加与更新。

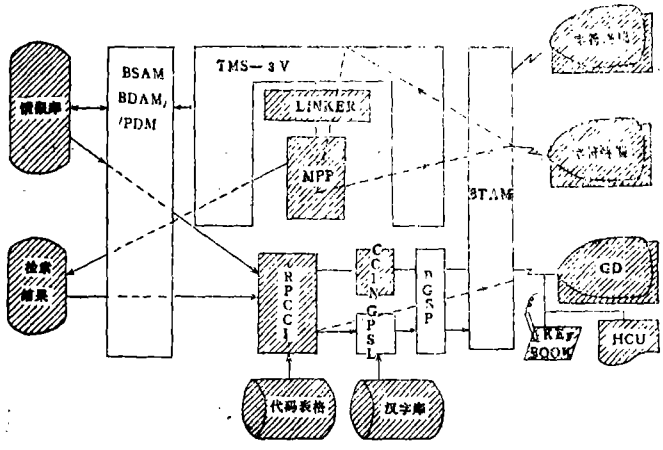


图6　实时联机环境中CCIRS结构方式

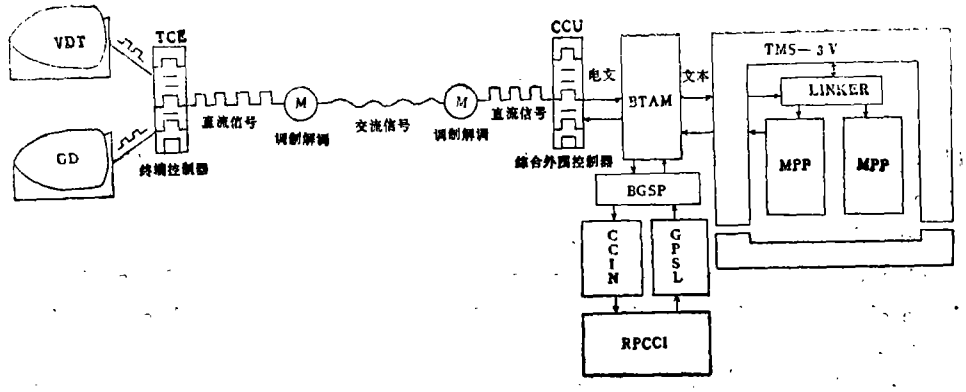


图7　实时环境中的信息传输与转换

图1至图7中：

VDT—Vedeo Data Terminal
GD—Graphi Display
TCE—Terminal Controllor
M—Modern
BTAM—Basic Telecommunication Access Method
TMS—Transaction Management Systems
BGSP—Basic Graphic Subroutine Package
GPSL—Graphic Plotting Subprogram Library
CCIP—Chinese Characters Iuput program
RPCCI—Retrieval Program of chinese Characters Information

参 考 资 料

〔1〕中华人民共和国国家标准：信息处理交换用的汉字字符集及其编碼，（1979）。
〔2〕日立制作所システム開発研究所，日本語情報検索システム，日立教育中心報告，(1980)，3。
〔3〕田中和明、武市宜之、稲川博之，カナ．キーワード入力による漢字情報の検索方式の提案，昭和55年度電子通訊学会総合全国大会論文集，1980年。
〔4〕日本情報処理学会，第21回全国大会プログラム，1980.5.21.
〔5〕HITAC：漢字処理KEIS (Kanji Processing Extended Information System) 概要，1980.5.9.
(6) HITAC：グラフィックサブルーチンパッケーヅ(BGSP)，昭和54年9月（第3版）。
〔7〕HITAC：Graphic Plotting Subprogram Library，昭和55年4月（第3版）。
〔8〕（日）中央電子株式会社応用システム部，日中友好技术交流会デモプログラム機能仕样書，1980.11.広州。

A Research Report on the Chinese Characters Information Retrieval System CCIRS (I)

Yao Qingda

Abstract

This paper is a research report on the Chinese characters information retrieval system called CCIRS.

During the designing of CCIRS, we suggested a scheme on the input Chinese characters, besides, we designed an on—line retrieval and processing system with Chinese characters information. The basic functions of CCIRS has been implemented on a M—150 computer.

On this part I, we introduce the functions and construction of CCIRS.