

微机数据库变长记录和压缩存储技术

刘瑞新 (黄河水利学校计算机研究室)

一、问题的提出

微机关系数据库系统 dBASE 或 FoxBASE+ 适合于对数值型、字符型等定长数据的处理,数据库结构一经建立,其字段宽度也就固定了。在设计数据库时,为了减少数据冗余,而又不使数据溢出,就要把字段宽度定义得不大不小,恰到好处。这种方法适用于数据宽度变化不大、稳定的场合。但是在很多情况下,如试题库、科技文摘、人事档案中的个人简历等等,其信息长度变化很大。这样的问题在实际工作中大量存在。如果利用 dBASE 或 FoxBASE+ 备注型字段(MEMO)来处理,每个备注数据都存放在以 512 字节为块的文件中,磁盘空间仍不适合处理变长信息。

本文正是为了解决这些问题,提出的一套处理变长信息的方法。

二、非压缩存储

在微机资源有限的情况下,设计系统应着重考虑使用方便、响应速度快和尽可能节省存储空间等问题。为了解决这三个问题,采用记录条数可变、使用记录指针和压缩数据信息相配合来实现处理变长信息的设计方法。为了叙述方便,下面以试题库系统为例来构造数据库。

1.主数据库结构

在信息系统中,多数项目信息长度比较固定,少数信息伸缩性大。为此在主库的结构设计中,对固定长度的信息采用定长字段,对变长信息采用链指针。下面为主库结构示意图,ZBK.DBF 是存储试题指标的指标库。

ZBK.DBF(主数据库)

字段名称	含义	数据类型	字段宽度
TH	题号:试题的唯一标识	C	5
TX	题型:试题的类型	N	2
TF	题分:试题的分数	N	2

2.子数据库结构

该试题库系统有两个子库,分别存储试题题文的内容(TWK.DBF) 和试题答案的内容(DWK.DBF),其数据库结构相同。

TWK.DBF(子数据库)

字段名称	含义	数据类型	字段宽度
TH3C	题号:试题的唯一标识	C	5
TEXT	试题题文内容	C	72

DWK.DBF(子数据库)

字段名称	含义	数据类型	字段宽度
TH	题号:试题的唯一标识	C	5
TEXT	试题题文内容	C	72

3.主库与子库的关系

将待处理的数据分为定长部分(数值型和较短的字符型)和变长部分,并对应地建立一个主数据库和若干个数据库。其中主库存储定长部分,子库存储变长部分。主库中的一条记录对应子库中的多条记录。

ZBK.DBF(主库)

TH	TX	TF
01001	2	5
01003	4	20
01008	3	10
...	...	...

TWK.DBF(子库)

TH	TEXT
01001	第一行试题
01001	第二行试题
01003	第一行试题
01008	第一行试题
01008	第一行试题
01008	第二行试题
...	...

DWK.DBF(子库)

TH	TEXT
01001	第一行答案
01003	第一行答案
01003	第二行答案
01008	第三行答案
01008	第一行答案
...	...

为了使主库记录与子库记录之间建立联系,主库和子库均定义了相同的关键字段 TH,即主库与子库通过 TH 相互联系。

4.程序的实现

通过变换,我们已把处理变长记录,变为处理定长记录。而对固定长度字段的处理,程序很好设计。以上面建立的数据库结构为基础,通过适当的程序控制,可以方便的实现主库与子库之间的联系。

三、压缩存储

1.数据库结构

采用上述处理方法,在一定程度上虽然大大降低了数据的冗余度,节省了存储空间,但经过我们对试题的分析、比较发现,试题的行数多为 2~5 行,一般行的行首、行中和行尾都在空格,每行一般情况下并不写满,尤其当试题中有公式或表达式时,每行后部都空着,白白浪费空间。因此在存储时,如果把这些空格、空行压缩掉,显示或打印时,再按原来的格式恢复。这样又可节省不小存储空间。如果要想实现上述功能,在存储时,就要记住每个空格、空行的位置及个数。这样势必提高程序的复杂程序、降低速度。在实际中我们发现,大多数情况下,空格出现在行尾,即不写满行,经过权衡,我们采用了比较简单的处理方法——压去尾部空格。经比较,采用压缩比不压缩约节省空间 1/5~1/3。

具体的实现方法是,某试题内容的编辑先在一个临时库中进行,该题编辑完后,在临时库中依次从第一条记录开始,压去尾部空格,加上一个行结束标志“~”,再压去第二条记录的尾部空格,与上一条连接起来。在压缩连接过程中,若连接后的字符串已够题文库字段的宽度,则存储到题文库中。直到把临时库中的所有记录都处理完。最后加一上个结束符“”(如果用索引关键字 TH 联系,则可不用结束符)。

采用压缩的方法,题文字段宽度的选取要适宜。如果直接采用前面子库的结构,TEXT 字段的宽度为 72 字节,则内容的最后一行平均要浪费 36 字节。因而可进一步优化数据库结构,减小字段宽度。压缩库的字段宽度可有如下变化:

字段宽度	1	2	3	4	6	8	12	18	24	35	72
迭加为 72 的次数	72	36	24	18	12	9	6	4	3	2	1

考虑同时兼顾节省存储空间(字段越小越省)和处理速度(字段越大越快),并参考一般试题长度(每行 72 个

字符,一般 2~5 行),选择适中的字段宽度。如选字段宽度为 24,每行 72 个字符要分为 3 个记录;若一题 5 行,则需要 15 个记录。

主库、子库和临时库的数据库结构为:

ZBK.DBF(主数据库)

字段名称	含义	数据类型	字段宽度
TH	题号:试题的唯一标识	C	5
TWH	本题在题文库中的开始记录号	N	4
DWH	本题在答案文库中的开始记录号	N	4

TEK.DBF(子数据库)

字段名称	含义	数据类型	字段宽度
TH	题号:试题的唯一标识	C	5
TEXT	试题题文内容	C	24

ZBK.DBF(主数据库)

字段名称	含义	数据类型	字段宽度
TH	题号:试题的唯一标识	C	5
TEXT	试题答案内容	C	24

TEMP.DBF(临时数据库)

字段名称	含义	数据类型	字段宽度
TEXT	试题内容	C	72

2.主库与子库的关系

主库(指标库)与子库(题文库、答案库)的联系,仍可通过相同的关键字(TH)联系,但由于记录数的增多,检索速度就会降低,因而采用链指针将更为迅速。

在主库中增加一个字段,存放相应子库中相应项目的第一个记录号。这样就可以用 GO <记录号>的方法,直接找到相应项目,速度很快,弥补了记录增多造成的速度减慢。但是考虑到在试题的编辑过程中,有可能出现死机、停电等非正常退出现象,这时可能仅存贮了一部分内容,指标库与题文库、答案库的联系就会被打断,题文和答案也可能没有存贮上结束符,将造成系统无法运行。为此题文库和答案库中仍保留了题号字段,这样在上述故障发生时,仍然可以通过题号来联系,增加了系统的安全,保证了系统的正常运行。

3.程序的实现

由于试题内容的录入仍在定长的临时库中进行,所以编程比较容易。内容录入完后,要压缩存储,所以要设计压缩程序。显示或打印时,要首先还原,所以要设计还原程序。

每录入一条主记录,还要把子记录的开始记录号存到主记录中。修改、删除子记录时,要同时调整主库中影响到的所在存储记录号的记录。

下面给出从临时库 TEMP.DBF 到题文库 TWK.DBF 的压缩转储程序,及显示题文库某题的还原程序。

ZBK.DBF(主库)

TWK.DBF(子库)

TH	TWH	DWH
01001	1	1
01003	3	2
01008	4	5
...	...	...

TH	TEXT
01001	第一行试题内容 1~第二行试
01001	题内容~第三行试题内容~
01003	第一行试题内容~
01008	第一行试题内容第一行试题
01008	内容第一行试题内容第一行
01008	试题内容~第二行内容~
...	...

* 临时库 TEMP.DBF 到题文库 TWK.DBF 的压缩转储程序

CHR1='~/

CHR2=' /

CELECT B &&题文库 TWK.DBF

APPEND BLANK

SELECT D 临时库 TEMP.DBF

NUMBER=RECCOUNT()

NUMBER1=1

GO TOP

DO WHILE NUMBER1 <= NUMBER

AA=TEXT

SELECT B

IF NUMBER1=NUMBER

MTEXT=TRIM(TEXT)+TRIM(AA)+CHR1+CHR2

ESSE

MTEXT=TRIM(TEXT)+TRIM(AA)+CHR1

ENDIF

GO WHILE.T.

REPLACE TEXT WITH SUBSTR(MTEXT,1,24), TH
WITH PTH

IF LEN(MTEXT)>24

MTEXT=SUBSTR(MTEXT,25)

APPEND BLANK

ELSE

EXIT

ENDIF

ENDDO

NUMBER1=NUMBER1+1

SELECT D

SKIP

ENDDO

RETURN

* 还原显示程序

SELECT B &&题文库 TWK.DBF

GO PTWH &&某题开始记录号

CHR1='~/

CHR2=' /

MTEXT=TEXT

GO WHILE.T.

GO WHILE AT(CHR1,MTEXT)=0.AND. AT (CHR2,
MTEST)=0

SKIP

MTEXT=MTEXT+TEXT

ENDDO

MSIT=AT(CHR1,MTEXT)

?SUBSTR(MTEXT,1,MSIT-1)

IF MSIT=LEN(MTEXT)

SKIP

MTEXT=TEXT

ELSE

MTEXT=SUBSTR(MTEXT,MSIT+1)

ENDIF

IF MTEXT=CHR2

EXIT

ENDIF

ENDDO

?

RETURN

我们在水利部中等专业学校通用试题库管理及试卷生成系统中,采用的就是这种技术。经三年多的使用,获得了满意的效果。