中国医学装备

# 基于疫情信息采集系统的汉字乱码技术改进*

于海铸① 王兴强②*

*基金项目：国家自然科学基金(81502896)“网状 meta 分析实效性评价方法及其广义线性混合效应模型的构建与应用研究”；山东省自然科学基金(BS2015YY022)“广义线性混合效应模型在网状 Meta 分析有效性评价中的应用研究”

①解放军联勤保障部队第 900 医院计算机应用与管理科 福建 福州 350025

②解放军联勤保障部队第 960 医院信息科 山东 济南 250031

**[摘要] 目的：**对疫情信息采集系统进行技术改进，以解决新型冠状病毒肺炎(COVID-19)疫情中信息采集系统读取身份证地址信息截取半个汉字造成乱码的问题。**方法：**通过修改数据库中表的字段长度，增加读取变量的内存大小，修改源程序数据窗口设置和替换字符截取函数，实现汉字的自动截取。通过定义数据库函数和结构化查询语言(SQL)语句，采用判断字符美国信息交换标准规范(ASCII)码的方法，处理历史乱码数据。**结果：**通过疫情信息采集系统技术改进，能够解决截取半个汉字造成乱码的问题，并能处理历史乱码数据。**结论：**疫情信息采集系统的技术改进，可解决汉字乱码，保障系统间的正常数据交互，效果良好，且实用性强。

**[关键词]** 新型冠状病毒肺炎(COVID-19)疫情；数据库；乱码；ASCII 码

*Solution Technology of Half Chinese Character Display Garbled Code in Epidemic Situation Information Collection System/YU Hai-zhu, WANG Xing–qiang//China Medical Equipment, 2020*

**[Abstract] Objective**: To solve the garbled code problem that epidemic situation information collection system read the address information of ID card intercepting half a Chinese character,the epidemic situation information collection system was improved. **Methods**: By modifying the field length of tables in database, increasing the size of read variable's memory, modifying the datawindow of source program and replacing the character capture function, the automatic capture of Chinese characters is realized. Through the establishment of SQL statements and database functions, the method of judging character ASCII code is adopted to deal with the historical garbled code data. **Results**: Through the technical improvement of the epidemic situation information collection system, the problem of garbled code caused by intercepting half a Chinese character can be solved, and the historical garbled data can be processed. **Conclusion**: The normal data interaction between the systems was guaranteed with improving the epidemic information collection system and solving the problem of garbled code. The method was characteristic with good effect and strong practicability.

**[Key words]** Novel coronavirus pneumonia (COVID-19); Database; Garbled code; ASCII code

**[First-author's address]** Department of Computer Applications and Management, The 900th Hospital of The PLA Joint Logistics Support Force, Fujian Fuzhou 350025, China.

根据抗击新型冠状病毒肺炎(COVID-19)疫情中采集每例患者和来院人员疫情相关信息的需求，解放军联勤保障部队第 900 医院紧急开发了一套疫情信息采集系统，该系统采集的数据可直接与医院信息系统(hospital information system，HIS)进行数据交互，为医生诊疗提供数据服务[1-3]。HIS 是医院的核心业务系统，基于 HIS 系统的二次开发可能会产生一定的安全隐患[4-6]。在实际应用中，HIS 数据库中保存的中文字符经常会遇到数据库乱码的问题[7-8]。HIS 数据库出现的汉字乱码会造成系统无法正常运行，绝大多数情况是因为数据库字符集不匹配而造成，因此可通过修改数据库字符集来解决[9]。然而，部分汉字被截取造成汉字乱码的情况，解决相对要复杂，需要根据具体情况来处理[10]。在开发疫情信息采集系统读取身份证地址信息时，截取半个汉字，造成数据乱码，致使无法正常与 HIS 进行数据交互。为此，本研究在剖析汉字乱码问题原因的基础上有针对性地进行技术改进，有效解决了汉字乱码问题，并对历史汉字乱码数据进行处理，保障了疫情信息采集系统的正常运行。

## 1 信息采集问题分析

医院在开发疫情信息采集系统时，采用计算机技术可扩展标记语言(extensible markup language，XML)标准字符串的格式与 HIS 进行数据交互[11-13]。在系统接口测试时，偶尔出现了“XML 根节点错误”的提示，错误的 XML 格式字符串局部数据为：“<?xml version=″1.0″ encoding=″gb2312″ standalone=″yes″?><PAT_INFO><MAILING_ADDRESS>某某市某某区无影山黄屯二区 6 号楼 1 单元 301?/MAILING_ADDRESS></PAT_INFO>”。

### 1.1 乱码数据

XML 标准对格式要求严格。“XML 根节点错误”的原因在于，节点 MAILING_ADDRESS 存在格式问题，其内容为采集人员的地址数据，而地址数据最后一个字符为乱码，该乱码与单字节字符“<”组合成了一个新的乱码字符，造成 XML 格式字符串无法正常解析，导致接口数据交互失败[14]。

对疫情信息采集系统数据库中采集记录表中的地址字段进行检索，发现最后一个字节为乱码的数据 1300 余条，均为详细的地址信息，格式为:“某省某市某区某小区几号楼几单元几室”，数据来源均为从身份证读取的地址信息。

### 1.2 乱码数据原因分析

(1)根据数据库采集记录表地址字段，可以分析找出数据库和源程序中造成汉字乱码的原因。

(2)数据库中该地址字段长度定义为 VARCHAR2(40 BYTE)，而乱码数据正好为 40 个字节，最后一个字符为乱码，前面的字符中存在半角字符。可以推测出＞40 个字节的地址数据被截取为 40 个字节，而汉字为全角字符，一个汉字占 2 个字节，获取数据时正好将 1 个汉字截取了一半，从而造成了汉字乱码。

(3)对源程序进行地址字段的检索，地址字段的来源为从数据库中读取、手工录入和从身份证中读取 3 个方面。排除前两个原因后，锁定了利用身份证读卡器读取身份证数据这个途径，发现身份证读卡器接口函数中在读取地址数据时，地址变量分配了 40 个字节的内存空间，造成了地址数据从源头上被截取。

## 2 信息采集系统方法设计与实现

针对上述原因，需要对数据库采集记录表地址字段长度进行扩充，对身份证读卡器接口函数地址变量内存大小进行扩充，此外，医院疫情信息采集系统采用 PowerBuilder 9.0 应用

程序开发环境进行开发，也需要修改相应程序设置和相关调用函数[15]。

### 2.1 修改数据库字段长度

修改数据库采集记录表 COLLECTION_MASTER_INDEX 中地址字段 MAILING_ADDRESS 的长度，结构化查询语言(structured query languege，SQL)语句如下：

```
ALTER TABLE  COLLECTION_MASTER_INDEX MODIFY ( MAILING_ADDRESS  VARCHAR2 (100 BYTE ) );
 COMMIT。
```

### 2.2 修改地址变量内存大小

修改身份证读卡器接口函数中地址变量内存大小：

```
string _address
//地址变量
_address = space(40)
//为变量分配内存，指定变量最大长度
ret = ss_id_query_address(handle, _address)
//读取地址信息
astu_id.address = _address
//身份证结构变量赋值
```

修改上述分配变量内存的函数参数即可，space(40)改为 space(128)。

### 2.3 修改源程序设置

修改数据库表结构字段长度后，源程序的数据窗口字段长度也需要相应修改，其方式有两种：①修改数据窗口该字段的 Edit 标签中 Limit 属性值，将原值 40 改为 100 即可；②在 SystemTree 列表中，打开数据窗口的 EditSource 右键菜单，直接修改数据窗口脚本，将 COLLECTION_MASTER_INDEX.MAILING_ADDRESS 列的 type 属性值，由原来的 char(40)改为 char(100)。同时源程序的数据窗口中 MAILING_ADDRESS 地址列的 Edit 标签中 AutoHorzScroll 属性值也要进行修改，设置为选中状态，否则数据窗口只取该地址列能显示部分的数据，过长的数据无法显示的部分则会自动截取。

PowerBuilder 9.0 在对字符串进行截取处理时，可以按字节个数和字符个数两种方式进行。Left()函数可以得到字符串左部指定字节个数的字符串，而 Leftw()函数可以得到字符串左部指定字符个数的字符串。在身份证读卡器接口函数获取地址数据后给数据窗口地址字段赋值时，需要将原来的 Left()函数改为 Leftw()函数，避免被截取。

## 3 信息采集系统历史数据处理

通过修改数据库表结构字段长度和源程序，可以实现截取半个汉字造成乱码问题的处理，但已经产生的历史数据仍需单独处理。对数据库采集记录表中 1300 余条地址乱码数据进行分析，发现最后一个字节均为乱码，只有 36 条数据的倒数第二个字符为汉字，其他均为数字。

### 3.1 乱码字节前为数字的情况处理

针对乱码前面为数字的情况，可以判断其采用美国信息交换标准规范(American standard code for Information Interchange，ASCII)码来解决。汉字和乱码字符的 ASCII 码＞128，而数字的 ASCII 码＜128，可以使用该条件作为过滤条件，执行如下 SQL 语句去掉地址数据的最后一个乱码字符：

```
UPDATE COLLECTION_MASTER_INDEX P
//更新数据库患者主索引表
SET P.MAILING_ADDRESS = SUBSTR(P.MAILING_ADDRESS, 0, LENGTH(P.MAILING_ADDRESS)
-1 )
//去掉地址数据的最后一个乱码字符
WHERE ASCII(SUBSTR(P.MAILING_ADDRESS,-1)) > 128
//乱码字符的 ASCII 码大于 128
AND ASCII (SUBSTR (SUBSTR (P.MAILING_ADDRESS,-2),1)) < 128 ;
//乱码前面数字的 ASCII 码小于 128
COMMIT;
//执行提交
```

其中 SUBSTR (P.MAILING_ADDRESS,-1)函数为取地址数据的最后一个字符,SUBSTR(SUBSTR (P.MAILING_ADDRESS,-2),1)函数为取地址数据的最后 2 个字符中的前一个。

**3.2 乱码字节前为汉字的情况处理**

针对乱码前面为汉字的情况，虽然数据量少，但是处理相对复杂，无法确定乱码前面有多少汉字或数字，也无法具体确定哪一个值的 ASCII 码。

经研究发现，身份证上的地址信息，符合公安机关关于住址书写规范的相关要求，其中单字节字符仅包含数字、大小写字母和个别符号，如“ - ”等。据此可以通过计算地址字符串的字节总数和单字节字符数量来判断地址字符串最后一个字节是否为乱码。

首先确定地址字符串的最后一个字节的 ASCII 码，是否＞128，如果大于，表明最后一个字符为汉字或为半个汉字。在此前提下，如果整个地址字符串的字节总数为偶数，而包含的单字节字符个数为奇数，可以判断出最后一个字符为乱码；如果整个地址字符串的字节总数为奇数，而包含的单字节字符个数为偶数，也可以判断最后一个字符为乱码。

设计判断地址字符串乱码函数如下：

```
CREATE OR REPLACE FUNCTION COMM.SF_JUDGECONFUSIONCODE (MAILSTR IN VARCHAR2)
  RETURN NUMBER IS
  RESULT   NUMBER;
  I   NUMBER(6);
  STR_NUM   NUMBER(6);  //单字节字符个数
  SUB_STR  VARCHAR2(4);
  STR_LEN   NUMBER(6);    //字符串字节个数
BEGIN
  I := 0;
  STR_NUM := 0;
  IF ASCII(SUBSTR(MAILSTR,-1)) < 128 THEN
//如果最后一个字节为单字节字符，判断为非乱码
    RETURN 0;
  END IF;
  STR_LEN := LENGTH(MAILSTR) ;
```

```
//字符串字节数
  LOOP
    I := I + 1;
    IF I > STR_LEN THEN
        EXIT;
    END IF;
    SUB_STR := SUBSTR (MAILSTR, I, 1) ;
IF ASCII (SUB_STR) = 45 OR (ASCII (SUB_STR) > 47 AND ASCII (SUB_STR) < 58 ) OR
(ASCII (SUB_STR) > 64 AND ASCII (SUB_STR) < 91 ) OR (ASCII (SUB_STR) > 96 AND ASCII
(SUB_STR) < 123 )  THEN
//“ - ”ASCII 码为 45，数字 ASCII 码为 48-57，大写字母 ASCII 码为 65-90，小写字
母 ASCII 码为 97-122
STR_NUM := STR_NUM + 1 ;
    END IF;
  END LOOP;
  IF (MOD(STR_LEN,2) = 1 AND MOD(STR_NUM,2) = 0) OR (MOD(STR_LEN,2) = 0 AND
MOD(STR_NUM,2) = 1) THEN
//字节总数为奇数且单字节字符个数为偶数,或字节总数为偶数且单字节字符个数为奇
数，判断为乱码
    RESULT := 1 ;
  ELSE
    RESULT := 0 ;
  END IF;

  RETURN RESULT;
END;
```

使用上述判断乱码函数，可以处理乱码字节前面为汉字的情况，处理 SQL 语句如下：

```
UPDATE COLLECTION_MASTER_INDEX P
//数据库患者主索引表
SET P.MAILING_ADDRESS = SUBSTR(P.MAILING_ADDRESS, 0, LENGTH(P.MAILING_ADDRESS)
-1 )
//去掉地址数据的最后一个乱码字符
WHERE COMM.SF_JUDGECONFUSIONCODE(P.MAILING_ADDRESS,-1) > 0 ;
//乱码判断函数
COMMIT;
//执行提交
```

此种处理方法同时也适用于乱码字节前面为数字或单字节字符的情况，但执行效率不如采用 3.1 中的方法效率高。

## 5 结论

在疫情信息采集系统技术改进中，通过修改疫情信息采集系统数据库中表的字段长度，增加读取变量的内存大小，修改源程序数据窗口设置和替换字符截取函数，解决了疫情信息采集系统读取身份证地址信息截取半个汉字显示乱码的问题。并通过建立 SQL 语句和数据库函数，采用判断字符 ASCII 码的方法，处理了历史乱码数据，避免了因数据乱码造成与 HIS 无法正常数据交互的问题，保障了新系统的正常上线运行，其使用效果良好，实用性强。

**参考文献**

[1]王洪岩.浅析医院信息系统(HIS)数据库的维护[J].中国新通信,2018,20(16):157.
[2]吴文俊,周彬,沈黎,等.医院信息系统 Sybase 数据库的维护管理[J].中国医疗设备,2015,30(8):81-83.
[3]冯海云,刘晓伟,李丹彤.基于 C/S 结构的分布式 HIS 架构的开发[J].医疗卫生装备,2017,38(5):66-69.
[4]谭跃庆,胡吉亭.药品通用名查询程序的开发与应用[J].中国医疗设备,2016,31(11):84-87.
[5]王兴强,刘长兴.HIS 中医保小目录警示功能的设计与实现[J].医疗卫生装备,2017,38(10):45-47.
[6]丁效军,郑理华,陈宇珂.基于 HIS 的医疗设备计量信息管理系统设计与开发[J].医疗卫生装备,2015,36(4):60-62.
[7]宫彦婷.医院信息系统 Oracle 数据库中导入数据中文乱码的解决技术[J].中国医学装备.2017,14(2):90-92.
[8]裴红云,谢庆,陈俊,等.ORACLE 地震前兆数据库汉字乱码解决方法探讨[J].高原地震.2017,29(4):45-48.
[9]张剑.ORACLE 字符集迁移及乱码问题的解析[J].赤子(上中旬),2016(22):133.
[10]于洪.基于 MySQL 数据库的 Java Web 开发中的中文乱码问题[J].信息与电脑(理论版),2015(16):10-11.
[11]咸玉龙,张景阳,陈鹏,等.基于 XML 的油气管道设备模型交互标准化方法[J].电子科学技术,2017,4(4):28-32.
[12]圣文顺,乔雨,邵琳洁.基于 XML 的异构数据库信息交互机制的实现[J].物联网技术,2019,9(12):32-35.
[13]潘凤,何志林.XML 异构数据接口在综合治超管理平台中的应用研究[J].山西师范大学学报(自然科学版),2018,32(4):31-35.
[14]王善发,吴道荣.dom4j 解析 xml 文档保存数据的乱码问题[J].保山学院学报,2016,35(5):63-67.
[15]蔡春情.基于 PowerBuilder 的医院信息系统优化[J].计算机时代,2019(10):67-69.

***通信作者：xingqiangwang@163.com**

**作者简介：于海铸，男，(1982- )，本科学历，助理工程师，从事医院信息化及数据库技术等研究工作。**