文章编号：1005-8451（2016）06-0043-03

.Net框架联合Oracle字符集问题研究

莫佩宏，吴紫薇
（中车长春轨道客车股份有限公司，长春 130062）

摘 要：本文通过大型国企软件应用中发现的汉字生僻字无法显示问题，从开发工具到系统配置及数据库连接方式等方面对问题发生的原因进行了全方位的剖析，最终找到以升级Oracle数据库字符集为最终解决方案，并对数据库字符集升级方式进行了全面测试，明确解决问题办法，并在实际工作中进行了实践，为企业级Oracle数据库的跨字符集升迁提供了良好的实践经验。

关键词：汉字；生僻字；配置；数据库；字符集
中图分类号：TP392 **文献标识码：**A

Analysis of Oracle character set problem under .Net frame

MO Peihong, WU Ziwei
(CRRC Changchun Railway Vehicles Co. Ltd., Changchun 130062, China)

Abstract: By software application in large state-owned enterprise, it was found that some rare used Chinese characters could not be displayed. In this text, the problem was comprehensively analyzed from aspects of tool development, system configuration, data base connection and so on. The solution was to upgrade Oracle database character set. The method of upgrading database character set was comprehensively tested to confirm the solution and check it in practice, which provided good practical experience in cross-character set upgrade of enterprise-level Oracle database.

Key words: Chinese character; rare used Chinese character; configuration; database; character set

中车长春轨道客车股份有限公司于 1992 开始使用 Oracle 数据库，应用系统早期开发工具一直使用 Oracle 产品提供的 Developer2000，随着网络技术的发展，ASP、JSP 等成为主要应用语言。最早自行开发的应用系统是人力资源管理系统，其中，各模块的开发语言主要使用 Developer2000，PB 和 ASP 等。

随着企业经营管理理念的不断提升和信息化技术的迅猛发展，2013 年，我们从全局出发，推行一体化概念。在这个思想的影响下，开始使用 UCML 企业级快速应用开发平台，该平台基于 .Net 框架、C# 语言，标准程度高，规范性强，软件开发方便快捷，开发效率大大提高。作为引导项目，选择了管理模式最为稳定的人力资源管理系统进行再次开发，数据库采用原有的 Oracle 数据库。

人力资源管理系统的数据大多与个人信息有关，表现最为明显的就是员工姓名，其中，涉及很多日常比较少见的生僻字。在此次人力资源系统整合升级的过程中发现，在显示员工姓名时，有些生僻字在系统中出现乱码，这个问题虽然不是很大，但是却直接影响着系统的使用性能，如果不能解决，会导致系统无法使用。

收稿日期：2015-12-02
作者简介：莫佩宏，教授级高级工程师；吴紫薇，高级工程师。

1 问题分析

1.1 从编程语言方面分析

1.1.1 配置编程语言中的字符编码

字符编码也称字集码，是把字符集中的字符编码为指定集合中某一对象，以便文本在计算机中存储和传递。常见的例子包括将拉丁字母表编码成摩斯电码和 ASCII。在使用的编程语言中，一般都会引用一套字符编码来解决文本转换问题。

我们使用的快速开发平台中配置文件中的字符编码为“UTF8”，此编码方式为目前主流的编码配置，可以根据不同的符号自动选择编码的长短。正常情况下可以满足大部分网页情况，但是，为了尝试解决汉字乱码问题，先后更换“GB2312”和“GBK”等专门针对简体中文字符的编码和扩展编码方式，都没有达到预期效果。

1.1.2 更换数据库连接方式

程序中数据库连接方式决定着不同的数据读取方式和转换方式，可能会导致特殊字符在转换过程中出现问题，因此在程序中尝试逐一更换几种目前主流的数据库连接方式。

（1）OracleClient 方式

引用类库：System.Data.OracleClient.dll。

命名空间：System.Data.OracleClient。

连接字符串：“data source=oratest;user id=scott;password=tiger”

（2）OleDb 方式

命名空间：System.Data.OleDb。

连接字符串：与 OracleClient 方式相比，要添加一个 provider，微软为“provider=MSDAORA.1;”或“provider=MSDAORA”，Oracle 为“provider='OraOleDb.Oracle';”。

（3）Oracle 提供的 Oracle Data Provider for .NET(ODP.net) 方式

引用类库：Oracle.DataAccesss.dll

命名空间：Oracle.DataAccess.Client 和 Oracle.DataAccess.Types

连接字符串：和 OleDb 方式格式相同，只是 provider 换为“Provider=OraOLEDB.Oracle.1”

结果表明，无法正确显示员工姓名中生僻字。

1.2 从框架方面分析

针对人力资源系统，从 2003 年开始，我们使用 ASP 网络语言开发了一系列应用模块，都没有发现有姓名生僻字无法显示的问题，而此次使用 .Net 框架却出现显示乱码问题，分析可能是由于数据读取解析方式不同造成的。

对比 ASP 这种前端解析开发语言，考虑 .Net 是否在框架方面有什么参数可配置，因此咨询微软支持工程师，对方表示到目前为止，微软方面没有这方面的配置。

1.3 从数据库方面分析

从 1992 年 Oracle6 版本开始引入应用，一直到现在，Oracle 数据库版本已经升级为 11 G，持续应用超过 20 年。虽然 Oracle 技术日新月异，但是由于引入的时间比较早，是 Oracle 针对中文配置字符集的初期，数据库字符集设置为 ZHS16CGB231280。

1.3.1 数据库客户端字符集配置一致性问题

联合数据库连接方式的改变，因为其中提到的前两种连接方式都是需要 .Net 程序发布服务器配置 Oracle Client 程序的，这就涉及到客户端的字符集是否与数据库服务器端字符集匹配的问题。将 .Net 程序发布服务器上的数据库客户端字符集通过注册表和环境变量（NLS_LANG）两种方式配置成与数据库服务器字符集相同，却依然无法正常显示生僻字。

1.3.2 数据库服务器字符集问题

数据库服务器字符集为 ZHS16CGB231280，查找相关使用手册，了解到该字符集只涵盖常用汉字 7 000 多个，一般的生僻字都是不包括在内的。在后期升级 Oracle 版本时考虑升级字符集，但是由于 Oracle 数据库的特殊性，其字符集修改必须满足新字符集是老字符集的超集（当一种字符集 A 的编码数值包含所有另一种字符集 B 的编码数值，并且两种字符集相同编码数值代表相同的字符时，则字符集 A 是字符集 B 的超级，或称字符集 B 是字符集 A 的子集）才可以修改，而后期出现的 ZHS16GBK 字符集虽囊括了近万的汉字，却无法保证是原字符集的超集，字符集的修改或转换有可能带来数据的丢失，存在很多不可预见的风险，因此在近 20 年的 Oracle 使用中，我们始终沿用了 ZHS16CGB231280 汉字字符集。

1.4 问题定位

综上所有测试结果，如果维持现有使用的工具和框架，唯一的解决方案就是针对数据库字符集进行升迁，将其字符集升迁至更大的汉字字符集，即 ZHS16GBK。

2 解决方案

2.1 安装配置新字符集下的Oracle数据库服务器

安装配置一台新的 Oracle 数据库服务器，在创建数据库时，可以指定字符集 (CHARACTER SET) 和国家字符集 (NATIONAL CHARACTER SET)。数据库字符集是指以什么编码格式用来存储 CHAR,VARCHAR2, CLOB, LONG 等类型数据，用来标示诸如表名、列名以及 PL/SQL 变量，存储 SQL 和 PL/SQL 程序单元等，新的服务器配置指定字符

（下转 P51）

时间为 max(I_1, I_2, …,I_k) =3.05 min；在“经济”策略下，两站之间共架设通过信号机 10 架，最大追踪间隔 max(I_1, I_2, …,I_k) =3.80 min<H (H=4 min)，满足追踪间隔要求。经检验，整体分布优化算法在两种策略下得出的闭塞分区划分方案均满足实际要求，说明算法有效。

4　结束语

本文将整体分布优化算法应用于闭塞分区的划分，分别在“效率”策略和“经济”策略下比较了整体分布优化算法和 PSO 算法的适应度值。整体分布优化算法搜索最优解的能力更强，鲁棒性也更优。最后通过实例进行仿真分析，并根据整体分布优化算法的输出结果形成布置方案。

参考文献：

[1] 王瑞峰．铁路信号运营基础 [M]. 北京：中国铁道出版社，2008.

[2] 康　宁，陈永刚，林俊婷，曹　岩．基于免疫粒子群算法的闭塞分区划分优化设计 [J]. 铁道标准设计，2013（11）.

[3] 刘剑锋，毛保华，侯忠生，等．基于遗传算法的区间自动闭塞信号机布置优化方法 [J]. 铁道学报，2006（8）.

[4] 左政伟，王思明．面向闭塞分区划分问题的模拟退火算法研究 [J]. 科学技术与工程，2012（12）.

[5] 林　祁．基于粒子群算法的铁路闭塞分区设计优化研究 [D]. 成都：西南交通大学，2009.

[6] 余炳辉．整体分布优化算法研究及应用 [M]. 成都：西南交通大学出版社，2012.

责任编辑　陈　蓉

（上接 P44）

集为 ZHS16GBK，几乎涵盖目前为止最多的汉字生僻字；国家字符集用于存储 NCHAR, NVARCHAR2, NCLOB 等类型数据，国家字符集实质上是为 Oracle 选择的附加字符集，主要作用是为了增强 Oracle 的字符处理能力，因为 NCHAR 数据类型可以提供对亚洲使用定长多字节编码的支持，而数据库字符集则不能。国家字符集在 Oracle9i 中进行了重新定义，只能在 unicode 编码中的 AF16UTF16 和 UTF8 中选择，默认值是 AF16UTF16，新的服务器配置选择默认值。

2.2　数据迁移

新整合的人力资源系统基于新的数据库服务器搭建，但原始数据需要无缺陷地进行迁移，迁移至新的数据库服务器。

2.2.1　Exp/Imp方式

（1）尝试在数据库服务器端使用 Exp 方式进行数据导出；（2）通过 Imp 方式进行数据导入，这种方式在导入导出期间做了字符集转换，但是通过对 Dmp 文件中的头文件分析，发现有数据丢失现象。在新导入的数据中，不但生僻字为乱码，同时正常的中文注释以及视图、存储过程等中存在的一些汉字代码也丢失变为乱码，导致视图和存储过程等失效。（3）我们使用客户端导入导出，导出时将客户端与源服务器端字符集配置一致，导入时使客户端与目标服务器端字符集配置一致，结果显示由于 Dmp 文件的头文件格式不同，无法实现导入过程。

2.2.2　数据泵Expdp/impdp方式

数据泵导出通常保存数据为与数据来源的数据库相同的字符集。数据泵导入时转换数据为目标数据库的字符集。在数据库会话启动之后，数据泵日志文件以 NLS_LANG 指定的语言写入。利用数据泵对用户进行数据导入后暂时没有发现其他问题，只是生僻字有显示乱码问题，通过客户端和电子表格等结合方式，在目标数据库中将生僻字进行统一修改，基本实现了数据迁移任务。

3　结束语

通过对人力资源管理系统进行再次开发，分析出现的问题及解决方案，我们可以总结如下：（1）针对 .Net 框架语言，由于其是微软开发的，使用 Oracle 数据库会存在一些不可预见的兼容问题；（2）对于 Oracle 而言，字符集的更改需要进行大量的测试工作，不断地发现问题和解决问题，最终实现字符集升迁和数据的成功迁移。

责任编辑　陈　蓉