维吾尔语在图书馆机读目录中的应用①

曹锦梅[1], 杨爱霞[2], 陈少鸿[3], 黄 杰[1]

[1](新疆医科大学, 乌鲁木齐 830054)
[2](新疆大学 外国语学院, 乌鲁木齐 830046)
[3](新疆会计干部培训中心, 乌鲁木齐 830002)

摘 要: 针对维吾尔语从图书馆角度、计算机、数据库角度进行分析, 解决维吾尔语在大型数据库 ORACLE 中的存储、显示和互联网检索的问题, 实现用民族语言检索机读目录, 从界面、提示、系统信息等方面实现民族语言本地化.

关键词: 维吾尔语; 图书馆; 机读目录

Uygur Language of Machine Readable Cataloging in Library

CAO Jin-Mei[1], YANG Ai-Xia[2], CHEN Shao-Hong[3], HUANG Jie[1]

[1](Xinjiang Medical University, Urumqi 830054, China)
[2](School of Foreign Languages, Xinjiang University, Urumqi 830046, China)
[3](Xinjiang Accounting training center, Urumqi 830002, China)

Abstract: Uygur is analyzed from the angles of the library, computer, database, solve the uygur language in a large database of ORACLE storage. This paper solves the displaying and the retrieval problems of the Internet, with national language retrieval machine readable catalog, from the aspects of interface, tips, system information such as national language localization.

Key words: Uygur; library; MARC

国外以美国最著名的国会图书馆, 采用的是与国内不同的 RLG 组织的维吾尔语罗马字母对照表对应的国外标准的维吾尔语新文字(与国内的不同)进行编目, 检索. 目前还没有实现维吾尔语(老文字)的机读目录的录入和检索. 国内按照教育部要求所有民族语言图书上同时只能用拉丁文转写[1]. 新疆民语委员会的发展计划中也指出, 民文维吾尔语、哈萨克语、柯尔克孜语的机读目录仍是空白. 在各民族地区图书馆都是和国家图书馆一样采用拉丁字母转写的方式录入到中文机读目录中去的. 只能用拉丁字母进行检索(实际是英文键盘字母组成的). 国内除蒙古语机读目录格式通过国家的验收, 其余民族文字机读目录都没有实现. 表 1 列出维吾尔文献目录的国内外现状.

表 1 国内外维吾尔文献目录数字鸿沟

对比方面	国外	国内
字符集	国外标准的维吾尔字符集	新疆本地使用的中国国家标准的维吾尔字符集
维吾尔语拉丁字母对照表	美国国会图书馆使用RLG(图书馆组织)标准的维吾尔语罗马化字母对照表[9]	新疆本地的维吾尔语新文字对照表
机读目录格式	MARC21	CNMARC(中文机读目录格式)
维吾尔语文献目录显示语种、可检索语种	有中文; 汉语拼音; 美国标准的维吾尔语罗马化的拉丁字母; 维吾尔语(阿拉伯字符)	只有新疆维吾尔新文字显示、检索.
语言字符集研发机构	维吾尔语计算机信息协会 (Uyghur Computer Science Association)(英国)研制维吾尔字库: UKIJ 系列	新疆大学信息科学与工程学院

用本民族的语言来查询图书馆数据库中本民族语言的文献资料应该是理所当然的事, 也是完全符合各民族读者的检索习惯的. 但目前新疆只有四十岁以上

①收稿时间:2014-03-14; 收到修改稿时间:2014-04-15

的维吾尔族学习过维吾尔新文字(拉丁字母), 年轻的一代学习的是维吾尔老文字(通行). 而在图书馆, 由于技术的原因, 没有办法用维吾尔老文字进行查询、检索, 这造成图书馆目录不能被读者完全利用, 达不到传播知识的目的. 尤其在这个创建和谐社会的时代, 研制支持民族文字的图书馆自动化系统, 具有非常重大社会意义, 多语种图书馆编目子系统更是重中之重. 如何实现民文信息在数据库中存储、查询和检索等处理及支持各种基于民文的数据库应用是一个重要问题. "[2]

1 图书馆编目软件关键技术

1.1 字符集的选择

GBK 是目前最常用的简体中文字符集. 国内大部分图书馆软件都支持 GBK 字符集. 那么 GBK 已经支持维吾尔文了, 为什么图书馆软件不能用 GB18030? 实际上, 维吾尔字符集的区位码的编码字形一直有争议[3], 直到 2008 年 1 月 1 日国家标准规范了维吾尔文、哈萨克文、柯尔克孜文编码字符集[4]. 发布在微软公司的网站(2008 年 5 月 13 日)(http://support.microsoft.com/kb/949793/zh-cn#appliesto)的系统补丁报告《应用程序不可能会发现区域设置或当用户使用 Unicode 的区域设置时, 不能启动》[5]中指出 Visual C++7.0 到 9.0 会出现问题, 其中就包括: 中国的蒙古语、中国的藏语、中国的彝语. 该文中没有提到中国的维吾尔语, 是因为维吾尔语采用的是 1256 阿拉伯语代码页. 为了解决这个问题, 于是产生了将 Unicode 编码规则和计算机的实际编码对应起来的一个规则, UTF 英文为 UCS Transformation Format, 即 UCS 转换格式, 目前常用的有 UTF-8、UTF-16、UTF-32 三种. 根据图书馆软件编目规则的要求, 只能采用 UTF-8 这种变长的 UNICODE 编码, 在字段标识符、字段代码按规定只能采用 ASCII 编码(UTF-8 的保留部分).

1.2 数据库系统地选择

由于图书馆编目系统机读目录格式变长的特点, 只能采用大型数据库, 我们选取 ORACLE[6], 没有采用微软 SQL2006 数据库, 因为图书馆软件对安全性要求高, 机读目录文件必须和书籍在历史时间上同在. ORACLE 数据库公司网站早就指出, 支持全球化语言[7]. 在数据库本地化设置(Oracle Locale Builder)的"其他项标签"中, 有书写方向的选项: 从左到右, 从右到左, 水平, 垂直四个可多选项, 默认是: 从左到右和水平. 所以, 在安装 ORACLE 数据库时, 没有采用通用默认的安装模式, 定制安装, 围绕多语种支持重新配置数据库运行参数.

1.3 运行模式的选择

在浏览器页面下使用 iSQLPLUS, 由于浏览器支持和数据库字符集相同的 UTF-8, 所以可以支持所有的符合规范的 Unicode 字符, 支持维吾尔语、哈萨克语、柯尔克孜语、藏语、蒙古语等, 只要是 UNICODE 编码的输入法都支持. 由于大多数软件开发工具针对国际化问题, 只提供了文字方向从左到右和从右到左的国际化设置, 还没有一种通用开发工具能实现从上到下的蒙古文和满语、达斡尔语、锡伯语文字的支持(不含专门用来文字竖排的转换软件). 但是客户服务器的软件运行模式响应速度快、及时, 开发的组件经过多年的积累远比 Web 模式多, 安全性更强. 本文采取如图 1 所示的运行模式.

(1)局域网: 通过应用服务器与数据库服务器连接的客户服务器模式, 负责整个系统的安全、配置、维护.

(2)广域网: 普通浏览器通过系统的应用服务器, 来读取查询数据库服务器的数据的 Web 开发模式.

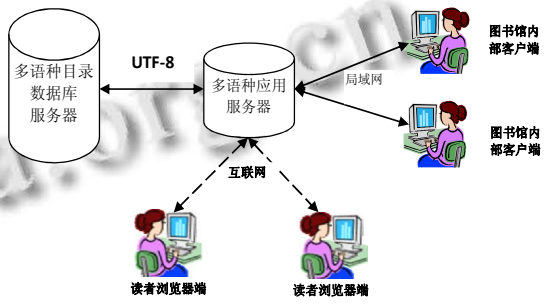

图 1 多语种图书馆编目系统网络模式

2 多语种图书馆编目软件分析

2.1 机读目录分析

数据库中字段存储的维吾尔语(阿拉伯语字符)是按逻辑顺序存储的, 机读目录的格式也是按照逻辑顺序存储, 子字段也是按字母大小写顺序从左到右逻辑存储. 这样无论是什么语种, 无论显示的方向是从左到右, 还是从右到左(维吾尔语、哈萨克语、柯尔克孜语), 都是逻辑顺序, 这样机读目录在多语种图书馆内部的业务规则各语种是统一的. 逻辑顺序的统一意味

着检索顺序的统一, 检索的前方一致性使我们不再为各个语种设计单独的检索策略, 又使多语种图书馆编目系统的设计简单化了.

2.2 解决民族文字本地化 WINDOWS 代码页与字库的问题

实际上, 代码页对应的字库中是没有维吾尔文字符的. UNICODE 独立代码页只有独立的藏语、蒙古语等字库文件. 而维吾尔语比较特殊, 它是和阿拉伯语共用代码页, 所以在阿拉伯语字库中, 包含了维吾尔语字符(新疆本地人因部分字符的形状不符合本地习惯, 很少使用), 所以微软专门提供了维吾尔字库(Microsoft Uighur, msuighur.ttf). 本文采用新疆大学研制的民族智能输入法, 包括了维吾尔语、哈萨克语、柯尔克孜语的多种字库, 但代码标识是相同的, 可以借鉴.

通过以上分析, 只要在数据库端和应用服务器端正确安装维吾尔语的 UNICODE 输入法就可以. 本次实验中, 下载的从 Vista 操作系统中移植出来的蒙藏彝维输入法, 安装到了 Windows XP 客户机上, 正常使用.

2.3 解决民族文字字库的选择问题

在 JAVA 开发环境中, 针对中文、维吾尔、藏文、蒙古文, 不宜直接指定使用任何汉字字库, 如宋体, 微软雅黑、仿宋, 否则只能显示汉字, 而不能显示民族文字(这实际上是国家标准没有真正执行). 本文安装了新疆大学维吾尔语智能输入法(与微软的输入法基本相同), 附带了一些字库, 进行测试成功.

3 多语种图书馆编目软件分析与实现

3.1 机读目录数据库平台

数据库和 J2EE 应用开发环境分别在两台计算机上. 在 Windows Vista 数据库服务器端监听 Windows XP 开发环境 Oracle Developer Suite10g, 应用服务器端的监听程序要自己重新设置, 源程序略. 数据库服务器端的防火墙要对开发环境(应用服务器)的 IP 地址、数据通讯的计算机端口 1521 开放, 才能正常使用.

平台: Win XP 端(Oracle Developer Suite10g 开发平台)和 Win Vista 端(ORACLE 库)

字符集: 数据库是 AL32UTF8 字符集和 Oracle Developer Suite10g 是 UTF-8 设置,

输入法: 安装在 Windows XP 上的从 Windows Vista 操作系统中移植出来的中国民族语言输入法(2008 年国内免费可以下载到的, 蒙古语、藏语、彝语、维吾尔语)和新疆大学的维吾尔语、哈萨克语、柯尔克孜语智能输入法.

测试结果: 普通 JSP 网页上各种民族文字都能正常传输、显示. 顺便编辑了一个最简单的不依赖数据库的纯 JSP 聊天室, 进行测试, 两台计算机联网都可正常输入、同步显示维、哈、柯、藏语等文字.

3.2 实现多语种编目系统的国际化

根据 ISO 国际标准化组织的语言代码 ISO639-1, 2, 3 的标准, 定义中文代码为 zh-CN, 维吾尔语代码为 ug-CN. (这里的语种代码是中国机读目录语种代码表的源头, 二者是继承关系, 直接选用 ISO639-3 标准三位是 iii, 二位是 ii).

3.3 实现多语种编目系统的本地化

采用在每个页面上添加语言选择按钮的方式, 让客户自定义选择界面语言(不是打开另一个页面, 只是及时将当前页面的语言的切换. 因为系统只有一个, 每个页面是所有语种共用的.

首先在 JSF 页面流控制配置文件 face-config.xml 中定义一个小的 Java 类, 用在管理豆(managed bean)中:

```
<managed-bean>
 <managed-bean-name>chenLocaleManager</managed-bean-name>
 <managed-bean-class>chenapp.chenLocaleManager</managed-bean-class>
 <managed-bean-scope>session</managed-bean-scope>
</managed-bean>
```

LocaleManager 类的源代码省略:

最后在每个需要本地化的页面上, 添加各语种的命令按钮, 维吾尔语切换按钮的源代码如下:

```
<af:commandButton
  text="uighur جۇڭخۇا خەلق) ئۇيغۇر يېزىقى جۇمھۇرىيىتى)"
actionListener="#{chenlocaleManager.chenchangeLocale}"
   id="chenchangeLocale_ug_CN"/>
```

这样测试成功, 实现了多语种数据库应用程序的语言切换问题, 完成了多语种机读目录的国际化、本地化的设置.

我们希望维吾尔语编目查询时符合阿拉伯语的习惯，从右到左的界面显示方式，包括上下滚动条在左面，这需要在 adf-faces-config.xml 添加 right-to-left 页面方向标签，判断读者浏览器的语言设置是否是维吾尔语(中国)，如果是就设置从右到左的属性为真.

<right-to-left>

#{view.locale.language in('ug_CN', 'kk_CN', 'kr_CN') ? 'true' : 'false'}

</right-to-left>

编写一个小小的类文件，如果从右到左的属性是真，传递参数给 jsp 页面，改变页面的左右方向的布局.

```
AdfFacesContext context =
AdfFacesContext.getCurrentInstance();
if (context.isRightToLeft())
{
  传递 RTL 从右到左的参数给页面的方向“dir”
}
```

IE8 以上支持模式，图 2 所示是从右到左的维吾尔语的机读目录查询页面：(由于是测试，截图中的各语种的书名、作者等文字是从各网页复制来的，内容可随意增改).

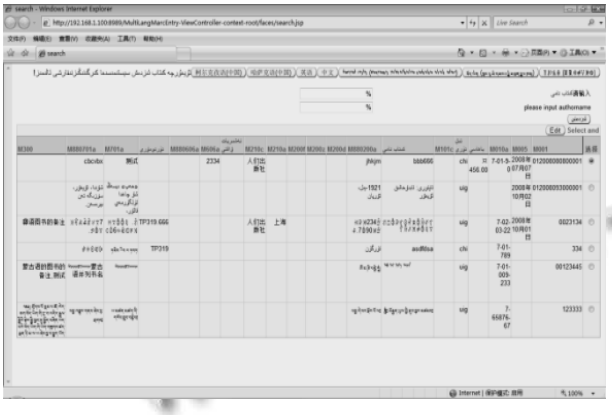

图 2 维吾尔语机读目录查询页面

4 总结

在 ORACLE ADF 和 JSF 技术的支持下，提出“统一”与“特色”的多语种编程思路，实现了维吾尔语和汉字共存于单一数据库中；统一的检索策略；统一互联网页面流程；统一用户界面，根据语言特性，呈现文字流方向的多样性. 开发民族语言的图书馆自动化系统，是实现民族团结、和谐社会的必然趋势，为民族文字数字图书馆的发展奠定坚实的基础.

参考文献

1 佟加·庆夫.新疆少数民族文字软件研发应用状况与发展建议.语言与翻译,2003,(1):72–76.

2 程伟,林河水,吴健,孙玉芳.数据库管理系统多民族语言支持研究.中文信息学报,2006,20(2):94–100.

3 鲍怀翘,金星华,宗成庆主编.少数民族语言信息技术研究进展—中国少数民族语言信息技术与语言资源建设学术研讨会论文集. 20040-04-11~12 日,北京.

4 新疆维吾尔自治区信息技术标准化委员会等. GB21669-2008.信息技术,维吾尔文 哈萨克文,柯尔克孜文编码字符集.北京:中国标准出版社,2008.

5 微软补丁报告.应用程序不可能会发现区域设置或当用户使用只 Unicode 的区域设置时,不能启动. http://support.microsoft.com/kb/949793/zh-cn#appliesto.[2008-05]

6 盖国强著循序渐进 Oracle:数据库管理、优化与备份恢复. Oracle 字符集.北京:人民邮电出版社,2007.

7 www.oracle.com. Globalization Support Oracle Unicode database support.甲骨文公司的白皮书,2005. http://www.oracle.com/technology/tech/globalization/pdf/TWP_AppDev_Unicode_10gR2.pdf

8 吾守尔·斯拉木,曹锦梅,朱雪莲,陈少鸿.维吾尔语、哈萨克语、柯尔克孜语在图书馆编目系统的应用.中文信息学报,2010,24(4):119–122.