

基于XML的高校网招录取数据交换技术的研究

白涛¹,王斌²,张太红¹,寇晓斌¹,冯向萍¹,王业²

(1. 新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830052;

2. 新疆农业大学 现代教育技术中心,新疆 乌鲁木齐 830052)

摘要:全国普通高校招生网上录取系统生成的投档单数据是各高校系统二次开发的数据源,高考政策的变化及系统数据结构、接口的改变会给高校系统带来巨大影响。基于可扩展标记语言(eXtensible Markup Language,XML)技术的异构数据交换模型主要包括XML生成模块、XML比对模块和XML数据合成模块。通过统一的方法对异构数据源进行结构化描述及字段关系映射,具有方便灵活及低成本的特点,是切实可行的方法。该模型已被用于新疆高等院校招生工作综合管理信息平台建设项目中,实现了自动化批量数据合成,实践证明是行之有效的。

关键词:高校招生;网上录取;可扩展标记语言;异构数据;数据交换

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2014)08-0202-04

doi:10.3969/j.issn.1673-629X.2014.08.048

Research on Data Exchange Technology for Universities Admissions Based on XML

BAI Tao¹,WANG Bin²,ZHANG Tai-hong¹,KOU Xiao-bin¹,FENG Xiang-ping¹,WANG Ye²

(1. College of Computer and Information Engineering,Xinjiang Agricultural University,

Urumqi 830052,China;

2. Modern Education Technology Center,Xinjiang Agricultural University,Urumqi 830052,China)

Abstract:The examinee electronic archives data created by nationwide general universities admissions on-line systems is the data source used for further development by universities themselves,the changes of college entrance examination policy and the alteration of data structure or interface will have a huge impact on the systems of universities. The heterogeneous data exchange model based on XML (eXtensible Markup Language) technology includes XML generator module,XML comparison module,XML data integration module. It describes the structure of heterogeneous data sources,and defines the mapping of data fields through a unified approach. It is convenient,flexible and low-cost,so it is the practical method. This model has been used in Xinjiang universities admissions integrated management information platform construction project and has realized the automation of bulk data combination. The practice has proved it is effective.

Key words:universities admissions;enrollment online;XML;heterogeneous data;data exchange

0 引言

目前,全国普通高校的招生录取工作是采用教育部统一下发的“全国普通高校招生网上录取系统”(以下简称“网招系统”),通过网络远程方式进行。录取过程中,考生的电子档案是高校审阅和录取的唯一依据^[1]。这些数据是高校录取结果查询、通知书发放、学籍注册及管理在校生信息数据库等的基础数据及来

源^[2]。高校通常都会根据自身需求研发系统与之对接,对网招系统导出的投档单数据进行进一步开发利用。但是近年来,随着我国高考改革进程的不断推进,各省区高考政策制度差异化趋势加大,网招系统升级导致的数据结构及接口发生变化,这些都要求高校系统能够以一种灵活、可靠、低成本方式来适应这种外部变化。因此,必须首先解决投档单数据源和本地数据源的异构数据交换问题,这将直接决定高校对这一重

收稿日期:2013-10-30

修回日期:2014-02-13

网络出版时间:2014-05-21

基金项目:新疆维吾尔自治区高校科研计划重点项目(XJEDU2011I24)

作者简介:白涛(1979-),男,甘肃兰州人,讲师,硕士,CCF会员,研究方向为数据库技术、信息检索;张太红,教授,博士,研究方向为数据库技术、农业信息化。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140524.2151.066.html>

要数据资源后期利用水平的高低。

1 问题描述

现有的网招系统为招生管理信息化提供了可靠的、标准化的生源数据来源,但同时也存在与各高校自身的信息系统相互独立,无法实现数据共享的缺陷^[3]。实际工作中,高校对投档单数据的利用至少要完成录取通知书、花名册打印和新生报到等基础功能,有些高校还希望进一步利用和挖掘录取数据资源,得到更有价值的信息作为决策依据。通过对网上录取基本信息的统计分析,对学校的生源质量、生源结构与报考率等做出科学的评价与汇总,为以后招生计划的投放、招生宣传工作的实施提供科学依据^[2]。以笔者所在的高校为例,除基本功能以外,还需要实现单科成绩统计(英语、数学)、录取成绩历史数据分析、生源地统计、第一志愿满足率等更进一步的功能。这些功能的实现必须首先解决以下几个问题:

- (1)高校录取生源范围包括全国各个省市自治区,每个省市自治区又将录取信息按批次、文理、民汉进行分类,致使每年要合成的数据批次众多、数据量大,人工单批次合成不仅效率低而且容易发生错误。
- (2)各省区高考政策制度差异化及网招系统更新、升级导致的数据接口及数据结构的变化,主要指投档单表结构字段的增减和字段长度的变化。
- (3)高考成绩项在网招系统中是自定义字段,各省结构不一致,导致成绩数据合成时容易发生错误,尤其是单科成绩。例如:2013 年安徽省高考成绩项共有 122 项,山东省仅为 10 项,“综合”成绩对应代码记录安徽为“04-综合”,而山东为“09-综合科成绩”,不仅单科成绩对应字段名不一致,而且代码内容字符串也不统一。
- (4)特殊省份的投档单数据需要特殊处理。例如广东省采用自己研发的招生录取系统,投档单结构与教育部网招系统的投档单结构不一致,数据格式也不统一;江苏省考试科目与分值标准与其他省份也不一致,这都需要特殊处理。

2 基于 XML 的数据交换模型

可扩展标记语言(eXtensible Markup Language, XML)是 SGML 的子集,提供了统一的方法进行结构化数据的描述和交换^[4-5]。XML 独立于应用平台和便于传输的特性为解决上述问题提供了便捷的手段^[6]。文中提出的基于 XML 的数据交换模型主要包括 XML 生成模块、XML 比对模块和 XML 数据合成模块,如图 1 所示。

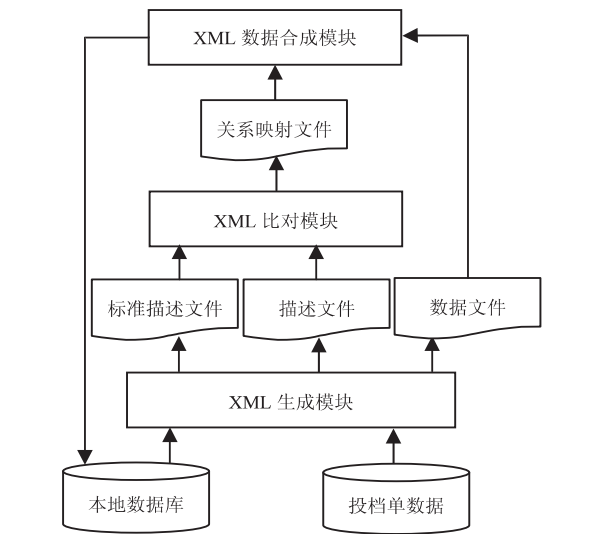


图 1 数据交换模型

2.1 XML 生成模块

投档单数据包括 6 个信息表和 18 个代码表, DBASE 3 格式,其中考生投档单表 T_TDD. dbf 存放了所有考生基本信息,是需要处理的最重要的数据表。通常高校自行开发的本地数据库只需要投档单数据源中的部分表,这取决于实际需求。XML 生成模块需要对用到的表生成 XML 格式的数据结构描述文件和数据文件^[7-8]。对本地数据库生成标准数据结构描述文件,这是进行表结构比对和数据合成的依据,第一次创建后,如果标准数据结构没有发生变化则无需重新生成。数据结构描述文件 XML 的节点包括 3 层:表节点、字段节点、字段描述节点,具体见表 1。

表 1 数据结构描述文件 XML 节点结构

类型	名称	属性	数量	说明
表节点	TableDef	Area, Class	1	表结构描述
字段节点	FieldDef	NO., Name	字段数	字段编号、名称
字段属性节点	DataType	—	字段数	数据类型
字段属性节点	Length	—	字段数	字段长度
字段属性节点	IsNULL	—	字段数	是否为空
字段属性节点	IsIndexField	—	字段数	是否为索引字段
字段属性节点	IsRequired	—	字段数	是否为必填字段

投档单数据文件 XML 的节点包括 2 层:表节点、记录节点,具体见表 2。

表 2 数据文件 XML 节点结构

类型	名称	属性	数量	说明
表节点	TableDef	Area, Class, Cnt	1	表结构描述
记录节点	RecordDef	所有字段数据	记录数	记录内容

2.2 XML 比对模块

XML 比对模块通过对 XML 生成模块创建的数据结构描述文件进行对比分析,完成对同名字段的兼容性检验以及建立非同名字段的映射关系^[9],最后生成 XML 格式的关系映射描述文件。该文件将作为数据

合成模块的基本依据。

对投档单数据源中用到的每一个数据表而言,需建立本地数据库中的同名表,通常情况下表结构的定义包含投档单数据表的全部或部分字段,此外还会包含一些自定义字段。以考生投档单表 T_TDD 为例,假定以 A 表示本地数据库中 T_TDD 的独有字段;B 表示本地数据库和投档单数据源中 T_TDD 的同名字段;C 表示投档单数据源中 T_TDD 的独有字段。XML 比对模块对 B 同名字段部分只需检查字段类型的兼容性即可,对 A 部分和 C 部分而言则需要按照某种规则建立映射关系。

2.3 XML 数据合成模块

XML 数据合成模块是首先读入投档单 XML 数据文件,对数据进行标准化处理,再根据 XML 比对模块创建的关系映射文件生成 SQL 语句,将数据合并至标准表^[10]。这里的标准化处理包括代码转换、专业名称标准化和考生号校验等工作。代码转换主要是各省区部分代码表并没有统一使用国家标准,需要进行转换;专业名称标准化主要是指专业名称后有时会带专业方向或加上“(定向)”等后缀,为了避免专业对应错误必须首先进行标准化;考生号校验主要针对广东省考生,广东省考生考号只有 10 位,而教育部系统考生号为 14 位,因此需要补齐前 4 位:“2 位年份+2 位地区编码”。

3 系统实现

文中采用 Delphi7 作为开发平台,SQL Server 2000 作为后台数据库,目标是把投档单数据源中需要的 DBASE 格式的数据表经过标准化后导入并合成到标准表中。为了实现数据库与 XML 的数据交互用到了 ADO 技术和 XML 编程接口 TXMLDocument。在上述的模型中,关键在于如何建立字段映射关系,因此 XML 比对模块的实现是难点。XML 生成模块和 XML 数据合成模块在成熟工具的支持下实现相对容易,限于篇幅不再详述,基本的两部分工作简要描述如下:

(1) 利用 ADOConnection 连接目标数据库,利用 TDataSet 的 TFields 类获取表结构信息和数据记录信息。主要用到的属性有 FieldNo、FieldName、DataType、Size、IsNull、IsIndexField 和 Required 等,其中,获取枚举类型 DataType 的值,需要用 GetEnumName 方法。函数原型如下:

```
function GetEnumName( TypeInfo: PTypeInfo; Value: Integer ): string;
```

(2) 利用 TXMLDocument 创建 XML 节点、设置节点属性,生成 XML 文件;或利用 XML Data Banding 方式访问已存在的 XML 文件,读取节点数据。

创建的数据结构描述文件如图 2 所示。

```
<?xml version="1.0" encoding="GB2312" ?>
- <TableDef Area="新疆" Class="汉语言">
+ <FieldDef NO="1" Name="ID">
- <FieldDef NO="2" Name="KSH">
  <DataType>ftWideString</DataType>
  <Length>14</Length>
  <IsNull>0</IsNull>
  <IsIndexField>0</IsIndexField>
  <IsRequired>0</IsRequired>
</FieldDef>
+ <FieldDef NO="3" Name="XH">
```

图 2 数据结构描述文件

XML 比对模块会对 XML 生成模块创建的数据结构描述文件进行比对和映射^[11-12]。如果生成的投档单表描述文件为 TDD.xml,SqlServer 中标准表名为 BZK,其表结构描述文件为 BZK.xml,则工作流程如图 3 所示。

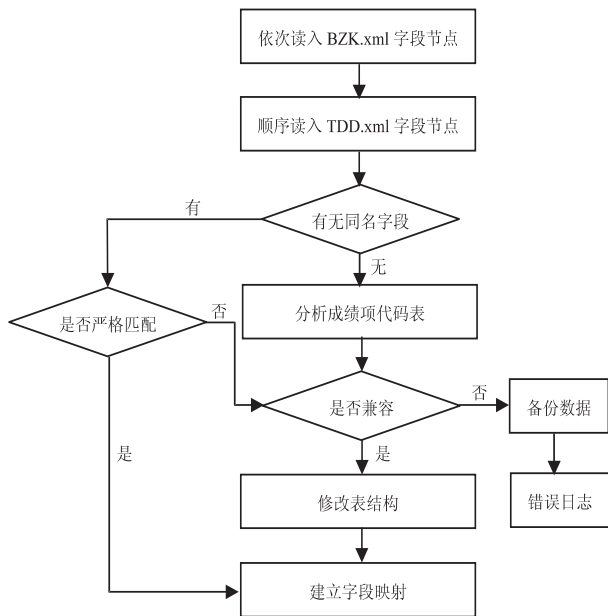


图 3 XML 比对模块流程图

在比对过程中逐一读取 BZK.xml 字段节点,顺序扫描 TDD.xml 的字段节点,查找同名字段。如果存在同名字段并严格匹配(字段属性均一致)则直接建立映射^[13];如果同名字段并不严格匹配,则需进一步判断是否兼容(BZK.xml 字段节点类型、大小是否包容 TDD.xml 对应字段节点),如果兼容或调整表结构后兼容则修改表结构,更新 BZK.xml,并建立映射;如果不兼容则记录错误日志,备份 TDD_Data.xml 数据文件。

需要重点说明的是,T_TDD 表结构中只有成绩项字段 Gkcjx01…GkcjxMN 是用户自定义的,即成绩项名称和字段名的对应关系是由各省区自行确定的,并不统一。而大多数高校对成绩数据的利用只需要语文、数学、英语和综合成绩,必须从多个成绩项字段中找出对应关系,这需要分析成绩项代码表 TD_CJXDM.dbf 中的记录,建立映射关系^[14],如图 4 所示。

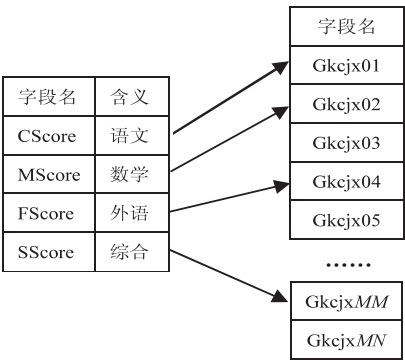


图 4 成绩项字段映射关系示意图

因为各省成绩项构成复杂,命名和顺序不统一,所以只能通过代码表记录进行关键字匹配,实践中为了提高准确性还应该结合成绩分析验证。由于目前高考单科成绩总分已知,一个投档单中成绩平均分应处于合理区间,如果字段对应错误,平均成绩会严重偏离。此外,由于总分是明确的,因此还可以将映射后的单科成绩相加与总分比对,如果不相等则表明字段对应发生错误,需要重新建立映射关系。

4 结束语

虽然高考网上招生录取工作模式已比较成熟,但随着高考改革进程的不断推进,各省区高考录取工作的差异性将会逐渐加大,网招系统还会不断更新和改进,这必然带来数据结构和接口模式的变化,即便是微小的改变也会对高校自身系统造成巨大影响。文中提出的基于 XML 的异构数据交换模型利用统一的方法对投档单数据进行结构化的描述和交换,具有实现简单、功能可靠、扩展性良好的特点,能够以较低的成本适应不断变化的使用环境。目前,该模型已被应用于新疆高校科研计划重点项目“新疆高等院校招生工作综合管理信息平台建设”中,效果良好。

参考文献:

[1] 邵庆辉. 一网当关:网上录取流程解析[J]. 招生考试通讯(高考版),2013(4):24-25.

[2] 张 强. 远程网上录取模式下高校招生工作初探[J]. 河南工业大学学报(社会科学版),2007,3(4):39-41.

[3] 陈 军,汪卫斌,王宏涛,等. 基于 B/S 与 C/S 结构的招生管理系统的实现[J]. 计算技术与自动化,2005,24(1):78-81.

[4] 颜廷良. 基于 XML 消息的安全数据交换平台研究与应用[J]. 计算机技术与发展,2013,23(2):173-176.

[5] 谢晓燕,王 浩,陈彦萍. 资源受限网络中高效 XML 交换的性能评估[J]. 计算机技术与发展,2013,23(6):54-58.

[6] 胡能发,唐为萍. 基于 XML 的通用异构数据交换模型[J]. 计算机工程与设计,2010,31(8):1743-1745.

[7] 赵 凯,赵正德. 低开销的异构数据交换[J]. 中国图象图形学报,2012,17(6):726-729.

[8] 刘 铮,刘 伟. XML 模式与关系模式间的映射冲突解决方法[J]. 计算机工程与设计,2010,31(17):3895-3898.

[9] 常 浩,安建成. 基于 XML 的异构数据交换模型的研究[J]. 电脑开发与应用,2011,24(3):27-29.

[10] 周红波,孙宇达,王继霞,等. 基于 XML 的数据交换及其参照完整性研究[J]. 计算机工程与设计,2006,27(14):2611-2613.

[11] 贾长云,朱跃龙,朱 敏. 基于 Huffman 编码与 XML 的大对象数据交换[J]. 计算机工程与应用,2006,42(19):177-179.

[12] 黄国言,郭 徽. 基于 XML 的协同设计中数据交换方法的研究[J]. 计算机工程与设计,2007,28(24):6000-6002.

[13] Tan Zijing, Zhang Liyong, Wang Wei, et al. XML data exchange with target constraints[J]. Information Processing & Management, 2013, 49(2):465-483.

[14] Seng Jia-Lang, Wong Zon. An intelligent XML-based multidimensional data cube exchange[J]. Expert Systems with Applications, 2012, 39(8):7371-7390.

(上接第 201 页)

[6] 孔令德. 计算机图形学基础教程[M]. 北京:清华大学出版社,2008.

[7] 彭 辉,刘善梅. 分形理论在植物形态模拟中的应用[J]. 农机化研究,2010,32(6):190-192.

[8] 汪富泉,罗朝盛. 基于迭代函数系统的分形算法及其应用[J]. 工程数学学报,2002,19(2):103-108.

[9] Lindenmayer A. Mathematical models for cellular interaction in development, Parts I and Parts II[J]. Journal of Theoretical Biology, 1968, 19:280-315.

[10] 邹运兰,杨志红,王仁芳. 基于分形几何的植物模拟研究[J]. 农机化研究,2012(1):195-198.

[11] 黄艳峰,薛占熬,陈 涛. 基于 L-系统的植物模拟研究[J]. 计算机工程与应用,2005,41(19):53-55.

[12] 丁 欢,万旺根,黄 炳,等. 三维嵌套 L 系统及其在植物模拟中的应用[J]. 计算机工程与应用,2009,45(5):207-209.

[13] Huang Yi, Sun Hai'an. Combining IFS and VQ in fractal image coding[J]. The Journal of China Universities of Posts and Telecommunications, 2000, 7(1):26-30.

[14] Zhou Jun, Chen Leiting, Liu Qihe, et al. Fractal-based 3D tree modeling[C]//Proc of 2010 international conference on computer design and applications. Qinhuangdao: [s. n.], 2010: 454-457.

基于XML的高校网招录取数据交换技术的研究

作者：[白涛](#)，[王斌](#)，[张太红](#)，[寇晓斌](#)，[冯向萍](#)，[王业](#)，[BAI Tao](#)，[WANG Bin](#)，[ZHANG Tai-hong](#)，[KOU Xiao-bin](#)，[FENG Xiang-ping](#)，[WANG Ye](#)

作者单位：[白涛](#)，[张太红](#)，[寇晓斌](#)，[冯向萍](#)，[BAI Tao](#)，[ZHANG Tai-hong](#)，[KOU Xiao-bin](#)，[FENG Xiang-ping](#)([新疆农业大学 计算机与信息工程学院](#)，[新疆 乌鲁木齐](#)，[830052](#))，[王斌](#)，[王业](#)，[WANG Bin](#)，[WANG Ye](#)([新疆农业大学 现代教育技术中心](#)，[新疆 乌鲁木齐](#)，[830052](#))

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(8)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjz201408048.aspx