一种批量信息检索方法及实现①

李 杰 宁 帆 赵国安 (北京邮电大学 网络教育学院 北京 100876)

摘 要: 在信息处理领域，存在大量数据信息需要进行校验，以确保数据的准确性、一致性、真实性，基于此提出了一种较为通用的批量数据检索方法，并开发了针对学位备案系统中学位授予信息的批量数据检索工具软件。该方法采用客户端的方案，使用数据库驱动程序以屏蔽异构数据源的差异，数据的读取、比对由批量数据检索软件执行；数据的检索基于批量检索方法定义的规则；数据库模式信息采用了数据库名-数据库表名-字段名组成的层次结构，以树来显示。开发的软件支持 Oracle,SQLserver,Excel 等数据源，并具有跨平台的特性。

关键词: 数据库；批量数据校验；Oracle；数据比对

Design and Realization of a Bulk Data Validation Method

LI Jie, NING Fan, ZHAO Guo-An

(Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: A large quantity of data need to be validated to ensure accuracy, consistency and authenticity in the field of information processing. This paper introduces a general method of bulk data validation and develops a tool application. This application is a client to the data source. It accesses the data sources through drivers provided by the individual data source vendor. So the application is independent of specific data source. The application reads and validates data. It retrieves data according to the rule defined by the method. Database schema information is shown in the form of tree, which contains the name of database, the name of table and the name of field. The application can validate data from Oracle, SQL Server, Excel and so on, and it can run cross-platform.

Keywords: database; bulk data validation; Oracle; data calibration

随着近年来社会信息化进程的快速发展，各个行业，领域都建立了大量的信息系统数据库。数据产生的及时性，收集的全面性已不再是信息采集工作的难题。如何充分利用系统存储的大量原始数据或派生数据成为摆在数据开发人员面前的新课题[1]。在采集了新的数据后，往往需要与数据库中的原有数据进行比对，以确保新采集数据的真实性和准确性。然而，通过 PB 或 SQL 访问数据库，专业水平要求高，非计算机专业的人员很难得心应手，给普通数据管理人员的工作设置了障碍。尤其是当需要检索，校验的数据很多时，其工作量无疑是巨大的。如果开发出一种批量校验工具软件来代替人工的校验，无疑将节省大量的劳动，并且提高信息校验的效率和准确性。

学位授予国家备案系统是教育主管部门为了及时了解学位授予及教育发展动态，对学位授予信息进行采集，汇总和整理的工作平台。其备案数据库存储大量的学位授予信息，并且每学期、每学年又有大量的学位授予信息存储到其中。另一方面，社会上大量的部门机构需要用自有学位信息与学位备案系统中的学1 位信息进行比对，以保证自有学位信息的真实性，一致性。经常的情况是，需要比对的数据量很大，例如，高校在研究生考试报名时，就需要对报考人员所需学历信息进行确定，以确保报考人员学历条件合格。目前，这些信息比对工作还停留在人工作业的阶段，

① 收稿时间:2009-08-19;收到修改稿时间:2009-09-23

迫切需要一种能方便、快捷、且准确性可靠的批量数据信息检索工具。

学位授予国家备案系统的后台备案数据库使用的是 Oracle 数据库，Oracle 数据库和 Microsoft Excel，是最适合不同用途的应用程序。Oracle 数据库可以处理大量数据，并且支持同时连接，无论性能和坚固性都是非常好的数据库服务器。另一方面，Excel 是适用于个人数据管理，具备各种各样的便利机能的客户端应用程序[2]，目前国内外很多教学，科研 ，设计等部门都在运用这一软件从事日常的数据表格处理，在国内，它也早已成为我们日常办公，学习的主要助手[3]。

依据访问数据源的位置可以分作基于服务器端方案和基于客户端方案[4]两大类。多数据库系统(multi-database)采用基于服务器端方案，这些方案有联邦数据库(federated database)，数据库透明网关(gateway)，中间件等[5]。如 DB2 的联邦数据库的方式，Oracle 的透明网关，SQL Server 的链接服务器及 DTS。它们主要提供跨数据库访问，但成本高，实现也比较复杂，不适合中小规模的应用[5]。基于客户端方案是采用数据库前段开发工具研制出批量数据校验工具软件，由该工具软件访问数据源并抽取数据。目前，国外专业 BI 厂商的产品已具有较强的批量数据校验的功能，而国内在这一领域还处在刚起步阶段，没有成熟的商业产品。另一方面，专业的 BI 厂商的产品价格昂贵，不适合中小企业使用。

本文作者提出了一种基于客户端的、批量数据检索方法，并以此开发了用于对学位信息批量检索的工具软件。该软件运行在 Windows 平台下，以 Excel 电子表格作为数据源；以 Oracle 作为数据库。它依靠数据访问接口连接数据源，根据批量检索方法进行批量检索。

1　批量数据检索方法

1.1 批量数据检索方法的设计思想

需要校验的一组数据(主动数据)存储在简单的 Excel 电子表格中。所谓简单的 Excel 表格，指的是类似关系数据库中的二维表，每一列代表一个数据字段，每一行代表一个数据记录，且不含复杂类型，如公式等的计算数据，表中可以有控制和显示格式存在，如字体格式，页面控制等[3]。例如学生学位信息 Excel 表，如图 1 所示，它包含的都是简单数据，就如同一张数据库的关系表，包含：姓名，出生年月，身份证件号码，学位证书编号四个字段。因为平时需要校验的数据都是比较简单的，这里先只考虑这种情况。

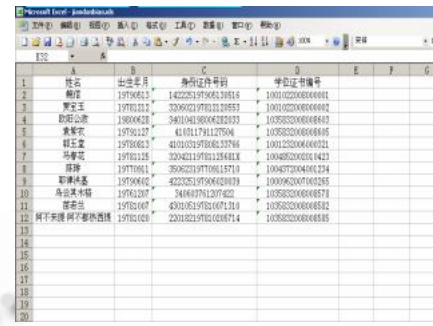

图 1　简单 Excel 表格实例图

由于需要检索的数据信息只是数据库中数据信息的部分而非全部数据，因此没必要采用数据仓库的方式复制数据库中的大批量数据。另外，批量检索对查询响应的时间要求比较高，采用中间件方式很难达到要求。该方法采用 ETL (Extract—Transform—Load. 抽取—转换—加载)技术流程思想，客户端在读取 Excel 中的数据信息后，按照统一格式建立一个临时的索引文档，以存储获得的 Excel 表格数据信息的相关描述信息，如行号，行数和列数，以及各个单元格的数据类型等。另一方面，在获取到 oracle 数据库的模式信息后，也以统一的索引文档格式存储到其对应的索引文档，其中包含有数据库名，数据库表名，记录序号，字段等信息。这些索引文档存储到索引文档库，为索引和查询提供数据接口。

在开始检索时，按照逐行逐列提取 Excel 表中数据与 Oracle 数据库中相应的表中每一条记录的相应的字段进行比对的方式。首先通过对 Excel 表格数据信息索引文档与数据库索引文档信息的比对，判断数据库表记录中的该字段类型是否与要比对的单元格的数据类型相对应，若是，则提取 Excel 中的相应单元格数据值与该字段进行比对；若是类型不相对应，则与下一字段进行同样的操作，直至与该条记录的某一字段比对成功或是比对完所有字段。如若 Excel 表格中某一行中的一个单元格的数据与 Oracle 数据库表中某一条记录的一个字段比对成功，则继续循环让 Excel 表格中该行的下一个单元格的数据继续与

Oracle 数据库表中的同一条记录中的其他字段进行比对；若是 Excel 表格中该行中的某一个单元格的数据没有与 Oracle 数据库表中的同一条记录中的其他字段比对成功，则循环与下一条记录的字段进行比对，循环操作，直至 Excel 表格中某行的所有单元格的数据都与 Oracle 数据库表中的同一条记录的不同字段比对成功或是比对结束。如果比对结束时，Excel 表格中某行的数据没有比对成功，则把该行数据存储在 Oracle 数据库的一个临时表中，以供显示来反馈用户。批量数据检索的流程图如图 2 所示。

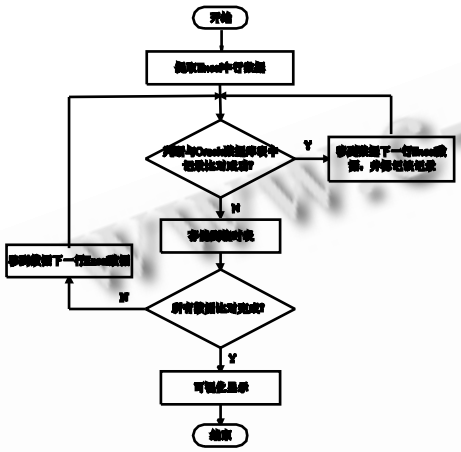

图 2　批量数据校验流程图

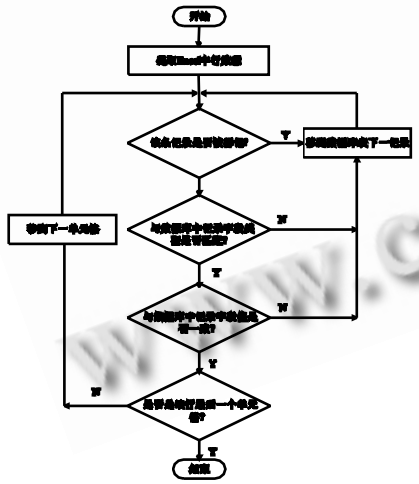

图 3　数据比对流程图

1.2 批量数据检索任务

批量数据检索任务的执行分成 6 步：

(1) 读取 Excel 表格中的数据信息，并建立一个临时索引文件以存储 Excel 表格中的数据信息的描述信息

(2) 获取 oracle 数据库的模式信息，并存储到其临时索引文件

(3) 读取 Excel 表格的索引文件，判断与数据库模式信息临时索引文件中相应数据库数据记录的字段的类型是否对应。如果类型对应则进行下面的步骤；不对应则与下一个字段进行比对。

(4) 根据每一个单元格的数据类型，执行 SQL 语句，将该单元格数据与 Oracle 数据库表中的记录的字段进行比对。比对成功则下移一个单元格，并返回到步骤(3)，不成功则与下一个字段进行比对。

(5) 将没有比对成功的 Excel 中的相应的数据信息(主动数据)存储到 Oracle 数据库的一个临时表中。

(6) 从 Oracle 数据库的临时表中提取数据，并显示以反馈用户。

可见，整个批量数据校验任务包含了执行数据校验工作所需的信息，包括数据源，数据源表，目标数据库以及目标数据库表等。

考虑到批量数据校验的复杂性，如果一次批量数据检索包含多个数据源表，或多个目标数据库表，则需把此次检索任务进一步的分解，使之只包含一个数据源和一个目标数据库表，以简化校验。

1.3 数据访问接口的选择

访问数据库管理系统(DBMS)必须选择数据访问接口。Windows 平台的数据访问接口很多，如 JDBC，ODBC，DAO，RDO，OleDB/ADO，ADO.NET 以及 BDE，等[5]。JDBC(Java Data Base Connectivity，java 数据库连接)是一种用于执行 SQL 语句的 Java API，可以为多种关系数据库提供统一访问，它由一组用 Java 语言编写的类和接口组成。JDBC 为工具/数据库开发人员提供了一个标准的 API，据此可以构建更高级的工具和接口，使数据库开发人员能够用纯 Java API 编写数据库应用程序[8]。这里采用 JDBC 数据接口，基于一下考虑：①JDBC API 与 ODBC 十分相似，有利于用户理解。②JDBC 使得编程人员从复杂的驱动器调用命令和函数中解脱出来，可以致力于应用程序中的关键地方。③JDBC 支持不同的关系数据库，使得程序的可移植性大大加强。④用户可以使用 JDBC-ODBC 桥驱动器将 JDBC 函数调用转换为 ODBC。⑤JDBC API 是面向对象的，可以让用户把常用的方法封装为一个类，以备后用。

1.4 对 Excel 表格数据的读取[6]

Java Excel 是一开放源码项目，通过它 Java 开发

人员可以读取 Excel 文件的内容、创建新的 Excel 文件、更新已经存在的 Excel 文件。Excel 中的 Excel 文件(工作簿)、工作表、行、列、单元格对应 Java Excel 中的 workbook、sheet、row、column 和 cell。读取 Excel 读取的流程为：①打开工作簿；②打开工作 Sheet 区；③找到某行某列的数据值。

一旦得到了 sheet 对象，就可以通过它来访问 Excel Cell(术语：单元格)。Java Excel 为 Cell 提供了一系列的方法来对 Cell 对象进行操作。如果仅仅是取得 Cell 的值，可以方便地通过 getContents()方法，它可以将任何类型的 Cell 值都作为一个字符串返回。如文本型，数字型，日期型等的 Cell 值，通过方法 getContents()，三种类型的返回值都是字符型。在得到 Cell 对象后，通过 getType()方法可以获得该单元格的类型，然后与 API 提供的基本类型相匹配，强制转换成相应的类型，最后调用相应的取值方法 getXXX()，就可以得到确定类型的值。API 提供了以下基本类型，与 Excel 的数据格式相对应，如图 4 所示：

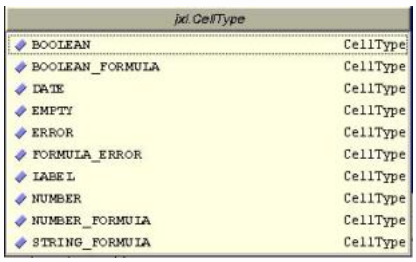

图 4　JDBC API 的基本数据类型图

1.5 Oracle 数据库模式信息的提取

一个关系数据库的设计主要包括关系模式(即基表结构)设计和完整性约束申明两部分[7]。基表结构定义了关系(表)的结构、属性(列)及其数据类型与长度等；完整性约束定义了语义施加在数据上的约束，包括关系层的全局约束、元组层的表约束以及属性层的列约束。这两部分信息构成了数据库模式的主要内容，并作为元数据存储于数据库的数据字典中[7]。

一般而言，数据库模式信息既可以直接从数据库的数据字典中提取，也可以通过分析 SQL 数据定义语言(DDL)语句来获得。然而，DDL 语句分析法存在诸多缺陷[7]。这里采用 JAVA API 来设计提取处理逻辑，通过读取数据字典来获取 Oracle 数据库模式信息。

JDBC API 中的包 java.sql 提供了用 java 语言访问和操纵 SQL 数据库的处理逻辑(接口和方法)。接口 DatabaseMetaData 中提供了获取数据库模式信息的许多方法。接口 DatabaseMetaData 的常用的提取模式信息使用的方法如表 1 所示。

表 1　提取模式信息常使用的方法

getTables ()　Search the description of	tables can be used in a given category
getTableTypes ()　Search Types of table s	can be used in this database
getColumns()　Search the description of	columns can be used in a given category
getDatabaseProduceName()　Acquire	the name of DBMS

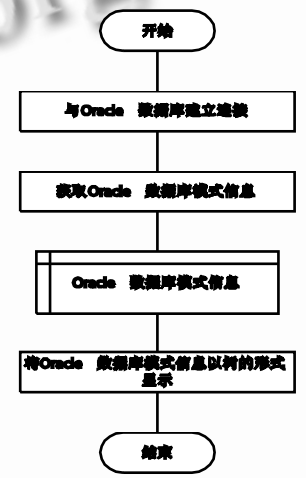

图 5　获取 Oracle 数据库模式信息流程图

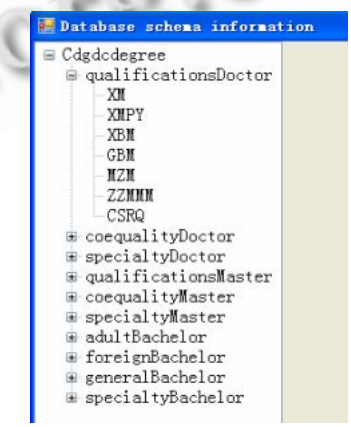

图 6　数据库模式信息的树状视图

JDBC API 与数据库建立连接后，可以获得一个 Connection 对象，通过调用 Connection 对象的 getMetaData()方法，即可获得一个 DatabaseMetaData 对象。从而获取到数据库的模式信息，并可以实

现以树的形式可视化显示“数据库名—表名—字段名”。组成的层次结构。获取 Oracle 数据库模式信息的流程图如图 5 所示。

连接到 oracle 数据库后，首先获取数据库的模式信息。然后以树的形式实现数据库名—数据库表名—表中的字段名组成的层次结构。用户可以方便的了解数据库的结构，以便选择相应的表作为批量检索的被动数据表。

数据库模式信息的树状视图如图 6 所示。

2 批量数据检索工具软件

采用面向对象的分析设计方法，以 Java 为程序设计语言，基于 Java2 Platform，Standard Edition，设计开发了该批量数据检索工具软件。它能跨平台运行，以图形用户界面提供用户操作，它通过 JDBC API 数据访问接口，连接到异构的数据源，能够支持 Oracle,SQL server，Excel,等数据源。批量数据检索工具软件系统体系结构如图 7 所示。

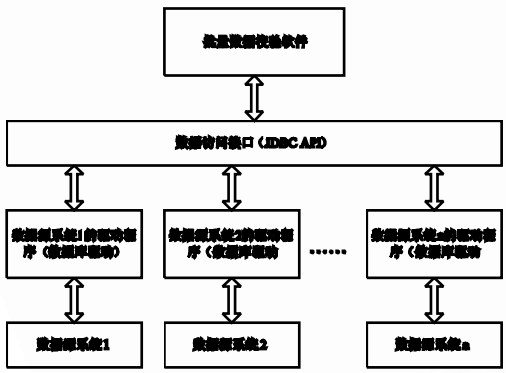

图 7 批量数据检索系统体系结构图

3 应用实例

教育部学位与研究生发展中心的网络版学位备案系统项目，后台数据库采用 Oracle 10g 数据库。其数据库具有存储的数据规模大、数据表及字段数量多、数据冗余等特点。

采用本批量校验工具软件，通过读取数十条以 Excel 保存的主动数据信息，与数据库中的被动数据进行批量校验，达到了设计的基本要求，能够满足批量校验的功能。

4 结论

从理论和技术的视角，通过研究分析，提出了一种较为通用的批量数据检索方法，该方法采用 JDBC API 数据访问接口及各自数据源驱动程序，屏蔽了异构数据源的差异，并可视化数据库模式信息。

采用面向对象的分析设计方法，设计实现了针对学位备案系统的学位授予信息进行批量数据校验的工具软件。该软件具有跨平台的特性，支持 Oracle,SQL server，Excel 等多数据源，成功地用于批量数据检索的工作中。研究表明，实现批量数据检索在技术上是可行的；该批量数据检索软件设计是合理的，实现是有效的。

参考文献

1 黄国莉. Excel 在提取 Oracle 数据中的应用.中国医院统计, 2006,13(3):265－266.

2 宋红,郭志刚,宋崴.将 Excel 数据导入至 Oracle 数据库的技术研究.佳木斯大学学报(自然科学版), 2006, 24(4):502－503.

3 邱宁. Excel 电子表格与数据库的数据转换.计算机应用与软件, 2004,21(10):24－25.

4 梁鹰,罗伟亮.异构数据库的数据转换在大型信息系统中的实现.计算机工程与应用, 2000,36(9):103－105.

5 刘如九,张振山,柴天佑.一种通用的多数据库间数据抽取方法及应用.北京交通大学学报, 2008,32(4): 14－17.

6 http://blog.chinaunix.net/u/22374/showart.php?id=154123, 2009,3.

7 许卓明,苏文萍.关系数据库模式信息的提取.河海大学学报, 2005,33(2):202－205.

8 魏永丰,刘立月.异构数据库中的 Oracle 与 SQL server 数据共享技术.华东交通大学学报, 2005,22(5):92－94.

9 JavaTM 2 Platform Standard Edition 5.0 API. http://gceclub.sun.com.cn/Java_Docs/html/zh_CN/api/, 2009,6.