

基于指令系统的 Excel 数据导入数据库方法研究

王允富, 武 尚, 张晓如

(江苏科技大学 计算机科学与工程学院, 江苏 镇江 212000)

摘 要: 针对现有 Excel 数据导入数据库方法通用性不够的问题, 提出并实现了一种基于指令系统的通用的数据导入方法. 设计了一套用于 Excel 数据抽取的指令系统, 给不同的 Excel 表一个可配置的指令文件, 通过指令解释器解析该配置文件, 便可自动完成 Excel 数据的抽取、及其数据向数据库的导入工作. 实践证明, 该套指令系统具有非常好的通用性.

关键词: Excel 文件; 指令系统; 数据抽取; 数据导入

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2014)09-0162-05

Excel Data Import into Database Analysis Based on Instruction System

WANG Yun-fu, WU Shang, ZHANG Xiao-ru

(School of Computer Science, Jiangsu University of Science And Technology, Zhenjiang 212000, China)

Abstract: A versatile data import method, based on instruction system, has been put forward to solve the problem of lack of versatility in methods for the Excel data into the database at present. We designed a set of instruction system for Excel data extraction, which provides different Excel table with a configurable command file that could be parsed by the command interpreter so as to finish Excel data extraction and data import into database automatically. Practical application shows that the instruction system has good versatility.

Key words: Excel file; instruction system; import data; extract data;

1 引言

由于 Excel 表格形式的多样性, 现有的 Excel 数据导入数据库的方法的通用性都不够^[1-3], 文献[4]中提出了一种知识引导的快速表格分割方法, 利用该方法可以对 Excel 进行分割, 分割之后根据知识进行处理, 但是构造一个完备的知识库及对知识库的维护是一件非常艰巨的任务; 文献[5]中利用 Excel 文件作为数据集成的中间件, 最终是将整个 Excel 文件作为一个数据源实现批量数据的导入数据库, 而不是以单元格为单位进行数据导入, 这种方法要求规范的 Excel 表格格式, 只有通过格式检查的文件才允许将其导入目标数据库; 文献[4]中提出了基于语法规则的通用的数据交换技术, 但是其局限性还是非常大的, 本文通过分析大量的不同形式

Excel 的特点, 设计出了一套提取 Excel 数据的指令系统, 且在实验过程中不断的得到完善和改进.

2 基于指令系统的 Excel 数据导入数据库方法

2.1 指令系统中用到的名词的定义

定义 1 主键属性名 (Primary attribute name) 即 Excel 表格中满足如下条件的任一属性名: 一是该属性值的属性名的数量要尽可能多; 二是该属性名的属性值和其他属性名的属性值能够形成一一对应关系, 不能有其他属性值无法与该主键属性值形成一一对应关系.

定义 2 指令系统即半结构化 Excel 数据抽取指令系统 (Semi-structured Excel Data Extract Instruction System) SEDEIS, 该指令系统的功能是从

不同格式的 Excel 表格中抽取所有的属性以及属性所对应的属性值。

定义 3 指令配置文件是指用来抽取具体的 Excel 表数据用到的一组指令集合^[5]。

定义 4 解释器:实现两种操作(Excel 表和数据库表)的互相沟通,需要通过了解二者关系的第三方(.NET 开发工具)来进行翻译解释,完成翻译工作的部件称之为解释器^[6]。

该系统中用到的指令解释器是指令配置文件和应用程序之间的一个软件组织,为应用程序从配置文件中解析出一组数据,并作为软件组织完成 Excel 数据的抽取及数据完成向数据库插入时使用的参数。

定义 5 空标识符即在给出的 Excel 表中,如果单元格中含有某特定字符,则表示该单元格的值为空,称该字符为空标识符。

定义 6 元组和指令名格式即指令的呈现形式,可表示为

instructionname(params string [] field, “;”);

其中, instructionname 表示指令的名称, field 表示元组的元素个数是可变的, “;” 冒号表示元组的分隔符即每个元组元素之间用冒号隔开, 小括号 “()” 表示元组的边界符。

定义 7 Excel 单元格可定义为: Cell = <row, column>, row 为 Excel 表格的行号, column 是表格的列号, 用尖括号括起来, 主要是用于和定义 6 中的圆括号区分; 行号和列号中间用逗号隔开, 用于和定义 6 中的字段分隔符冒号 (“;”) 区分; 这样可以简化指令的解析。

定义 8 分隔符函数(DelimitedFunction)即在抽取 Excel 数据时, 如果单元格内含有分隔符时, 需要调用该函数对单元格内容进行分割, 该函数作为指令的一个字段, 该函数可以形式化的表示为

Spilt(params string [] sign).

其含义为根据分隔符对单元格 < i, j > 进行分割的函数, params string [] sign 表示的是一个可变的参数, 此参数分为几种情况:

(1) 当该参数为 (“sign”) 时, 该函数表示为 Spilt (“sign”), 那么该指令表达的意义是首先取出单元格 < i, j > 单元格的值 value, 然后按照分隔符 “sign” 对 value 进行分割, 把取得的值依次作为与属性名相对应的属性值。

(2) 当该参数为 (“sign”, “num”) 时, 该函数表示为 Spilt (“sign”, “num”), 那么该指令表达的意

义是首先把字符串型 num 转换为整型, 其次取出单元格 < i, j > 单元格的值 value, 然后根据分隔符 “sign” 对 value 进行分割, 取得第 num 个字段的值作为与属性名相对应的属性值。

(3) 当参数为 (“sign1”, num1, “sign2”, num2) 时, 该函数表示为 Spilt (“sign1”, num1, “sign2”, num2), 那么该指令表达的意义是首先把字符串型的参数 num1 和 num2 转换成整型变量; 其次, 取出单元格 < i, j > 单元格的值 value1; 再次, 根据分隔符 “sign1” 对 value1 进行分割, 取得第 num1 个字段的值 value2; 最后根据分隔符 sign2 对 value2 进行分割, 取得第 num2 个字段值作为与属性名相对应的属性值。

(4) 如果分割符的个数多于两个的时候, 就按照以上格式写出 Spilt(params string [] sign) 函数的参数。

定义 9 连接符是指把多个单元格中的内容拼接起来, 整体作为一个属性名的属性值, 用符号 “+” 作为连接符。

2.2 指令系统(SEDEIS)的使用守则

为了指令解释器的实现更简单、快捷, 规定把指令写入配置文件的时候要严格遵守以下规则。

(1) 遵守 SEDEIS 的元组和指令名格式, 否则会解析出错。

(2) 第一条指令一定是地址指令, 即首先给出数据源所在的地址。

(3) 第二条指令一定是关系表指令, 即给出用来存储抽取的 Excel 数据的关系表的表名称以及该表的序列名称。

(4) 第三条指令一定是空指令, 如果数据源中没有空标识符, 则第三条指令一般是抽取主键属性名属性值的指令; 如果不存在主键属性, 则可以是其他的指令的任何一条指令。

(5) 除地址指令、关系表指令、空指令(含有空标识符的情况下)要求顺序以外, 其他指令都不分先后顺序。

2.3 基于指令系统的 Excel 数据导入数据库方法流程

该方法由指令系统、指令解释器、ExcelDataExtraction 类、GenerateSql 类四部分组成。用户首先要理解指令系统中每条指令的含义以及使用方法, 然后写出具体表格需要哪些指令语句, 最后根据指令的使用守则, 把相关指令写入指令配置文件; 接着解释器会自动对指令配置文件进行解析, 得到一组指

令中的各个字段值传递给 ExcelDataExtraction 类中方法使用,返回大量的 Excel 表中的数据以及创建表格的表名等参数;这些数据和参数会传递给 GenerateSql 类中的方法,由这些方法自动生成创建数据库表、sql 语句的生成、以及执行 sql 语句完成数据的插入工作.过程如图 2 所示.

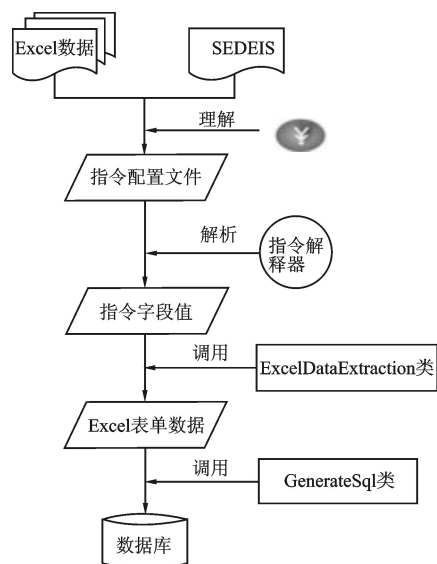


图 1 Excel 数据导入数据库的流程

3 Excel 数据抽取指令系统的设计方法

3.1 Excel 数据抽取指令系统的设计要素及分析

在 Excel 数据抽取指令系统的设计过程中,主要考虑以下几个因素.

(1) 指令系统的完备性,即指令丰富、功能齐全、使用方便,应具有所有基本指令.

(2) 指令系统的性能用完成对 Excel 数据抽取时所用的指令条数来衡量,指令条数越少,则数据抽取的速度越快,性能越好.

(3) 指令系统的复杂性,即指令系统不能太复杂,要做到结构简单,容易理解,易于使用以及利于解析.

(4) 指令系统要具有可扩展性.当指令系统的指令无法满足用户需要的时候,用户可以需要自定义一些指令,完成特定的数据抽取功能.

上述因素是存在相互制约的,不可能同时达到最优.因此,在该指令系统的设计过程中,需要根据设计目标和约束条件,综合考虑各种因素.由于 Excel 数据抽取指令系统的设计目标是能够让用户使用起来简单,并完成尽可能多的形式的 Excel 的数据抽取任务,因此,在上述设计要素中要优先考虑指令系统的完备性、复杂性等^[7].

3.2 Excel 数据抽取指令系统的设计方法

3.2.1 Excel 数据抽取指令系统在设计过程中应遵循的规则

为了使指令结构简单、易于理解、方便解释器的解析,在指令设计的过程中应严格按照如下规则:

(1) 在结构上采用元组和指令名的形式.

(2) 在指令名的命名上要遵循一定的命名规范.

3.2.2 Excel 数据抽取指令系统中指令介绍

根据上述分析,Excel 数据抽取指令的结构采用元组和指令名的格式,共十种指令格式,其中指令名、分隔符都是字符串类型,其形态如下所示.

(1) 空指令格式

Empty("sign")

其中,sign 表示如果单元格的内容是 sign,则说明此单元格为空.

(2) 地址指令格式

Address(srcpath;num)

其中,srcpath 表示源 Excel 表格的地址,其类型为字符串型;num 表示要取得该 Excel 表格 Workbook 的第 num 个 Worksheet,其类型为整型.

(3) 关系表指令

RelationTable(tablename;sequencename)

其中,tablename 表示存储已抽取 Excel 数据的关系表的表名称,其类型为字符串型;sequencename 表示的是表 tablename 的序列名称,序列是为了实现表的主键的自增功能,其类型为字符串型.

(4) 取值指令格式

Values(attributename;<i,j>;type)

其中,attributename 表示属性名,<i,j>表示 i 行 j 列单元格,该指令表达的意义是取出 i 行 j 列单元格值,作为属性名 attributename 的属性值;type 表示属性值的数据类型.

(5) 带分隔符的取值指令格式

DelimitedValues(attributename;<i,j>;type;Spilt(params string [] sign))

其中,attributename 表示属性名,<i,j>表示 i 行 j 列单元格,type 表示的是单元格内的值的数据类型,Spilt(params string [] sign)表示的是分隔符函数.

(6) 区域取值指令格式

DomainValues(attributedename;<i,j>;<m,m>;type)

其中,DomainValues 表示指令名称,attributedename 表示主键属性名,<i,j>表示 i 行 j 列单元

格, $\langle m, m \rangle$ 表示最大行最大列单元格. 该指令表达的意义是把从 i 行 j 列开始到最大行最大列范围的所有单元格内的值依次取出来, 全部作为主键属性名 $attributedename$ 的属性值, $type$ 表示所有属性值的数据类型.

(7) 带分隔符的区域取值指令格式

`DelimitedDomainValues (attributedename;  $\langle i, j \rangle$ ;  $\langle m, m \rangle$ ; type;`

`Spilt(params string [] sign))`

其中, `DomainValues` 表示指令名称, `attributedename` 表示主键属性名, $\langle i, j \rangle$ 表示 i 行 j 列单元格, $\langle m, m \rangle$ 表示最大行最大列单元格; `type` 表示区域内的属性值的数据类型. `Spilt (params string [] sign)` 表示的是根据分隔符对每个单元格 $\langle x, y \rangle$ 进行分割的函数, `Spilt (params string [] sign)` 表示的是分隔符函数.

(8) 映射指令格式

`Correspondence (primaryattributedename;  $\langle i, j \rangle$ ; attributedename;  $\langle \text{num}, j \rangle$ ; type)`

其中 `Correspondence` 表示指令名称, `primaryattributedename` 表示主键属性名, $\langle i, j \rangle$ 表示 i 行 j 列单元格, 该单元格内的属性值的数据类型是已知的, `attributedename` 表示的是与 $\langle i, j \rangle$ 单元格的值相对应的属性名, $\langle \text{num}, j \rangle$ 表示 num 行 j 列单元格, 该单元格内的属性值的数据类型为 `type`. 该指令表达的意义是与 $\langle i, j \rangle$ 单元格的值相对应的属性名 `attributedename` 的属性值是单元格 $\langle \text{num}, j \rangle$ 的值.

(9) 带分隔符的映射指令格式

`DelimitedCorrespondence (primaryattributedename;  $\langle i, j \rangle$ ; attributedename;  $\langle \text{num1}, j \rangle$ ;`

`type; Spilt(params string [] sign))`

其中, `Delimitedcorrespondence` 表示指令名称, `primaryattributedename` 表示主键属性名, $\langle i, j \rangle$ 表示 i 行 j 列单元格, `attributedename` 表示的是与 $\langle i, j \rangle$ 单元格的值相对应的属性名, $\langle \text{num1}, j \rangle$ 表示 num 行 j 列单元格, 单元格内的属性值的数据类型为 `type`; `Spilt (params string [] sign)` 表示分隔符函数.

(10) 带连接符的映射指令格式

`AttachedCorrespondence (primaryattributename;  $\langle i, j \rangle$ ; attributename;  $\langle \text{num1}, j \rangle$ ; “+”;  $\langle \text{num2}, j \rangle$ ; type)`

其中, `AttachedCorrespondence` 表示指令名称, `primaryattributename` 表示主键属性名, $\langle i, j \rangle$ 表示 i 行 j 列单元格, `attributename` 表示的是与 $\langle i, j \rangle$ 单元格的值相对应的属性名, $\langle \text{num1}, j \rangle$ 表示 num1 行 j 列单元格, “+”表示连接符, $\langle \text{num2}, j \rangle$

表示 num2 行 j 列单元格. 该指令表达的意义是与主键属性名 `primaryattributename` 的属性值对应的另一个属性名 `attributename` 的属性值是把单元格 $\langle \text{num1}, j \rangle$ 和 $\langle \text{num2}, j \rangle$ 内的值拼接成串, 该属性值串的数据类型为 `type`.

3.3 Excel 数据抽取指令系统的使用示例

通过将图 1 中的数据插入数据库, 来说明指令系统的使用方法:

(1) 图 1 中的表的指令配置文件如下:

```
Address(C:\Users\user\Desktop\hospital.xlsx;2)
RelationTable(hospital;seq_hospital)
Empty(“—”)
Delimitedomainvalues(expertname;
<4,2>;<m,m>;Spilt(“\n”))
Correspondence(expertname;<i,j>;
professionaltitle;<3,j>)
Correspondence(expertname;<i,j>;office;<i,1>)
AttachedCorrespondence (expertname; <i,j>; dia-
loguetime;<1,j>;“+”;<2,j>)
```

(2) 然后把配置文件的地址传递给指令解释器, 其将自动完成指令的解析, 得到一组参数 `Params`; 由于指令的格式固定, 解析起来比较容易.

(3) 把 `Params` 作为参数传递给 `ExcelDataExtraction` 类中的函数, 抽取 Excel 表格中的数据, 并存储在相应的数据结构 `DS` 中.

(4) 根据存储在 `DS` 中的数据, 调用 `GenerateSql` 类中的函数, 生成 SQL 语句, 完成数据向数据库的插入工作. 数据库关系表截图如图 2 所示.

4 结束语

指令系统对 Excel 数据的抽取是单元格级别的抽取, 甚至达到了单元格中带有分隔符的部分内容的抽取; 其类似于手工抽取 Excel 数据, 它告诉后续模块去抽取 Excel 表单中那个区域中的数据、数据类型是什么、按照何种条件去抽取数据, 然后加载到指定的数据库的关系表中. 与手工不同的是, 其中的指令配置文件, 将指令独立于数据的抽取过程, 使数据的抽取过程具有高度的通用性. 此外, 较现有的方法而言, 文中考虑的问题更为全面, 符合人的思维习惯, 更容易理解. 同时在指令格式的设计的过程中, 充分考虑了元组形式容易解析的特点, 具有很好的实用性, 减轻了编程的复杂程度. 该指令系统实现了对不同形式的 Excel 的数据进行快速抽取及自动插入到目标数据库中, 并利用大量的不同领域、不同表结构的 Excel 表数据对该指令系统进行验证, 证明了该指令系统不仅易于理解和使

ID	EXPERTNAME	PROFITLE	OFFICE	DIALOGUETIME
1	张古田	副主任医师	泌尿外科	星期一上午
2	燕翔	副主任医师	泌尿外科	星期一上午
3	连惠波	副主任医师	泌尿外科	星期一下午
4	甘卫东	主任医师	泌尿外科	星期二上午、星期二下午
5	张士伟	副主任医师	泌尿外科	星期二上午、星期二下午
6	王斌	主任医师	骨科	星期一上午、星期二下午
7	邱勇	主任医师	骨科	星期一下午
8	骆东山	主任医师	骨科	星期一下午
9	王骏飞	副主任医师	骨科	星期一下午
10	朱锋	副主任医师	骨科	星期一下午
11	朱丽华	主任医师	骨科	星期二上午
12	朱泽章	主任医师	骨科	星期二上午
13	任杰	副主任医师	骨科	星期二上午、星期二下午
14	蒋青	主任医师	骨科	星期二上午
15	袁毅路	主任医师	普外科	星期一上午
16	尹卫民	副主任医师	普外科	星期一上午
17	包善华	副主任医师	普外科	星期一上午
18	徐庆祥	主任医师	普外科	星期一下午
19	桑剑锋	副主任医师	普外科	星期一下午
20	陈蓉	副主任医师	普外科	星期一下午
21	管文贤	主任医师	普外科	星期二上午
22	周铁	副主任医师	普外科	星期二上午、星期二下午
23	陈刚	主任医师	普外科	星期二下午
24	张双玲	副主任医师	普外科	星期二下午

图2 数据库关系表截图

用,而且灵活性很强及具有可扩展性,具有很强的实用价值.不足之处:一是由于一张 Excel 表格对应一个配置文件导入到一个数据库关系表中,存在数据的冗余问题,如图2的科室,坐诊时间;二是没有考虑如何抽取多张相关联的 Excel 表数据.

(上接第161页)

调度的过程中,无法避免由于调度任务量过大造成数据节点冲突,从而导致调度效率降低的弊端.为此,提出一种基于数据节点冲突避免的云计算环境下海量数据高效调度方法.将整个调度任务划分成多个子任务,对全部调度任务中的子任务量进行计算,为模型的建立提供准确的依据.根据目标数据在数据节点的传输情况,对数据节点的任务量进行计算,根据计算结果,建立云计算环境下海量数据调度模型.实验结果表明,利用改进算法进行云计算环境下海量数据调度,能够有效提高调度的效率,避免了数据节点发生冲突的情况,取得了令人满意的效果.

参考文献:

- [1] 丁警,陈国良,顾钧. 计算网格环境下一个统一的资源映射策略[J]. 软件学报,2002,13(7):1303-1308.
- [2] 吴惠中等. 基于遗传算法的网格资源调度算法[J]. 计算机研究与发展,2004,41(12):2195-2199.
- [3] 汪德帅,张一川,张斌. 支持多租约 SaaS 应用按需服务的负载均衡策略[J]. 沈阳:东北大学学报:自然科学

参考文献:

- [1] 王晓刚,杨春金. Excel 到 SQL Server 数据库的数据导入导出技术研究[J]. 计算机与数字工程,2007,35(3):58-61.
- [2] 孙雪梅,丁军航,原明亭. 基于 C#.NET 的 Excel 表格数据导入数据库技术研究[J]. 信息技术与信息化,2009(6):109-110.
- [3] 王喆,潘懋,郭艳军. 基于 XML 元数据和 Schema 的 Excel 信息提取研究[J]. 计算机工程与应用,2008,(33):135-137.
- [4] 王湘波,王家华. Excel 与数据库间基于语法规则的数据交换技术[J]. 微电子学与计算机,2008,25(1):37-41.
- [5] 潘兆勇. 在 NET 中自定义配置文件[J]. 中文信息,2003(8):43-50.
- [6] 谭涛,陈汶滨,陈亚丽,等. NET 远程控制 Excel 在储量管理系统的应用研究[J]. 计算机应用与软件,2007,(02):43-44+71.
- [7] 曲英杰,刘卫东,战嘉瑾. 可重构密码协处理器指令系统的设计方法[J]. 计算机工程与应用,2004,(2):56-59.

作者简介:

- 王允富 男,(1986-),硕士研究生. 研究方向为智能信息处理.
武尚 男,(1989-),硕士研究生. 研究方向为智能信息处理.
张晓如 女,(1963-),教授. 研究方向为智能信息处理.

学版,2011,32(3):352-355.

- [4] 余永红,柏文阳. 强制数据隐私和用户隐私的外包数据库服务研究[J]. 计算机应用研究,2011,28(1):304-308.
- [5] 尚明生. 网格计算中的任务调度模型研究[J]. 计算机工程,2006,32(2):7-9.
- [6] 钟求喜,陈火旺. 任务分配与调度中遗传算子的设计[J]. 国防科技大学学报,2000,22(3):34-38.
- [7] 周文俊,曹健. 基于预测及蚁群算法的云计算资源调度策略[J]. 计算机仿真,2012,09(29):239-242.
- [8] 赵春燕. 云环境下作业调度算法研究与实现[D]. 北京:北京交通大学,2009:5-31.
- [8] 宋远骏,杨孝宗,李德毅,崔东华. 多机多任务实时系统云调度策略[J]. 计算机学报,2000,23(10):1107-1110.
- [10] 刘燕. 基于云计算信息处理系统体系结构设计[J]. 科技通报,2012,28(18):100-102.

作者简介:

- 方晖 男,(1976-),硕士,工程师. 研究方向为计算机软件技术、教育技术、实验室管理.