◎数据库、信号与信息处理◎

基于 XML 元数据和 Schema 的 Excel 信息提取研究

王 喆,潘 懋,郭艳军
WANG Zhe,PAN Mao,GUO Yan-jun

北京大学 地球与空间科学学院,北京 100871
School of Earth and Space Sciences,Peking University,Beijing 100871,China
E-mail:sonichappy@gmail.com

WANG Zhe,PAN Mao,GUO Yan-jun.Excel information extraction based on XML metadata and schema.Computer Engineering and Applications,2008,44(33):135-137.

Abstract: For the dump difficulty from complex Excel report forms using by mining enterprises and databases,bring forward the use of XML as the metadata model language to describe the information in Excel report.In addition,the use of XML Schema definition of XML metadata syntax,making applications based on meta data to extract information in Excel reports become more flexible and general.

Key words: Extensible Markup Language(XML);excel;schema;metadata

摘 要:针对地质工作单位中使用的复杂格式的 Excel 报表中的信息同数据库之间转储的困难,提出了采用 XML 语言描述的元数据模型来辅助表达 Excel 报表中的信息。使用 XML Schema 定义 XML 元数据语法,使得基于元数据的应用程序来提取 Excel 报表中信息变得更加灵活通用。

关键词:可扩展标记语言(XML);Excel;Schema;元数据
DOI:10.3778/j.issn.1002-8331.2008.33.042 **文章编号:**1002-8331(2008)33-0135-03 **文献标识码:**A **中图分类号:**TP302.1

1 引言

计算机技术的发展为社会带来了巨大的变革,使得人们的信息记录方式从传统的纸张过渡到了电子化时代。Microsoft 公司的 Excel 电子表格软件成为企事业单位中记录表格化数据的首选。Excel 在数据运算、数据加工、图表表达、数据分析统计等方面有优势,同时又简单易用。但与数据库系统相比,Excel 在数据冗余、数据共享、数据检索、数据关系表达、数据独立性、数据完整性、数据安全性、数据恢复、并发控制等方面表现不足。数据库系统专业性太强,一般工作人员很难直接操作它,因此经常遇到需要在 Excel 与后台数据库系统中交换数据的情形。在前端,工作人员仍方便地使用 Excel 制作电子表格;在后台,数据一般都保存在数据库系统中,以方便查询和统计[1]。本文介绍了使用 XML(Extensible Markup Language,可扩展标记语言)描述 Excel 报表格式的元数据信息,并基于该元数据构造存储过程的方法,使得能够通过通用的导入程序实现 Excel 数据到数据库的转储。同时讨论了该元数据 XML 文档的 Schema 定义。

2 Excel 信息提取方案

2.1 常用的提取方法

在数据量非常大的情况下,逐个将 Excel 中的数据录入到数据库既费时又费力,又很难保证正确性,这样的代价是无法承受的。有些数据库系统会自带数据导入模块,例如:SQLServer 2000,能够将格式比较良好的 Excel 文件自动化导入到数据库中;但是,该类型模块的使用局限性比较大,通常情况下只适用于行列非常规范的 Excel 报表文件,如图 1(a)所示;

	A	B	C	D
1	编号	姓名	年龄	住址
2	1	张三	24	北京市朝阳区望京小区
3	2	李四	23	北京市海淀区万柳小区
4	3	赵五	25	北京市崇文区北纬路小区

(a)

	A	B	C	D	E	F	G
1	钻孔类型		探矿孔			统一编号	
2	矿区名称		黑石砬子铁矿区			工程号	ZK417
3	坐标m	X	孔口	4545828.968			
4			孔底	4545562.97			
5		Y	孔口	41500497.519			
6			孔底	41500520.520			
7		H	孔口	77.719			
8			孔底	-806.124			

(b)

图 1 Excel 报表示例

基金项目:北京市多参数立体地质调查项目(the Multi-parameter 3D Geological Survey of Beijing under Grant No.200313000045)。
作者简介:王喆(1981-),博士生,主要研究方向:空间数据库;潘懋(1954-),教授,博士生导师,主要研究方向:信息地质、构造地质等;郭艳军(1980-),博士生,主要研究方向:三维地理信息系统。
收稿日期:2008-05-07 **修回日期:**2008-06-04

此外,本身就能导入 Excel 数据的数据库也不多见,仅限于 Microsoft 公司出品的 SQLServer 和 Access 产品,适用范围不大。

对于图 1(b)所示的这种格式比较复杂的 Excel 报表的导入策略通常是根据具体问题编写相应的导入程序来实现;虽然解决了问题,但是该方案的缺点也很明显:导入程序结构复杂、扩展性差、通用性不够等。如何让数据交换方法具有通用性成为亟待解决得问题。

2.2 XML 元数据方案

2.2.1 XML 元数据

元数据用于描述数据的内容、覆盖范围、质量、管理方式、数据的所有者、数据的提供方式等信息,是数据与数据用户之间的桥梁; 元数据可以独立于数据出现。使用元数据来描述 Excel 文件中的信息能够适用于各种复杂情况,其容易扩展、易解读、为异构数据的交换带来了方便。

XML 是由 W3C(World Wide Web Consortium,万维网联盟)提出应用于 Internet 环境中的数据交换技术,它以清晰的层次化结构、良好的可扩展性、自描述性、开放性等特点引起了各领域的关注,并得到推广采纳[2]。利用 XML 作为元数据的描述语言结合了 XML 自身的特点以及元数据的重要意义, 其优越性和作用十分巨大。

通过上面对元数据特点的描述,将元数据引入到 Excel 数据到数据库的转储流程中需要解决的问题如下:

(1)引入元数据使导入程序能与报表的具体格式相互独立,即使报表中数据的位置、合并格式等经常变动也不会影响程序的使用;

(2)使导入程序能与数据库的具体设计相互独立,即使数据库的结构设计发生变化,也不能影响报表的导入。

2.2.2 元数据的使用

由上面的分析可知, 在 Excel 报表导入数据库的过程中, 首先要对每个类型的 Excel 报表定制一个 XML 元数据文件, 使得在导入模块和 Excel 之间形成一个语义层。导入程序首先读取 XML 元数据信息, 并从元数据中定位获取报表中需要存入数据库的信息;另外,报表中信息同数据库表之间的对应关系也从元数据中获取,为后面的入库做准备。采用这种方式,导入程序就完全从 Excel 报表和数据库的具体设计中分离了出来,使其具有较强的通用性。整个导入流程如图 2 所示。

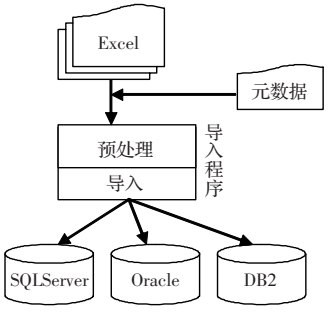

图 2 Excel 报表导入数据库流程

3 基于 XML Schema 的元数据模型

3.1 基于 XML 的 Excel 信息描述

3.1.1 root、group 元素与分块描述

XML 语法的规定,每个 XML 文件都是由一系列彼此之间形成树形结构关系的元素来构成的, 并且有且只有一个根元素。使用 root 作为每个 XML 元数据文件的根元素,其他的所有元素都是 root 子孙元素。

论文采用了元数据分块描述 Excel 报表的方法,此法可将单个 Excel 报表中的信息映射到不同的数据库表。用 group 元素标识 Excel 报表的一个分块,并作为 root 元素的子元素。因此,每一个 root 将包含一个或者多个 group 元素。

3.1.2 dbInfo 和 cell 元素

分块元素 group 包含一个 dbInfo 元素和一个或多个 cell 元素。dbInfo 元素和 cell 元素分别记录了该分块对应的数据库表信息和分块中每个单元格的信息。同数据库表之间建立映射关系,通常采用记录数据库表名称及其每个单元格对应的字段名称的方法, 在导入程序中根据这些信息来构造数据库 insert 语句并执行。但是这种方法在具体应用的时候仍然会比较死板,内部实现需要处理很多细节。考虑到使用导入程序的用户多采用企业型数据库,如 Oracle、SQLServer、DB2 等,论文采用了一种更加简洁高效的方式来表达报表同数据库的映射方法:在 dbInfo 中记录存储该分块信息所使用的存储过程名称,在每个 cell 中记录从该单元格提取的数据作为存储过程参数的位置。因为存储过程的代码是独立于导入程序的,还可以在数据插入之前或之后做一些处理, 而这些处理对导入程序是屏蔽的,提高了系统的灵活性。另外,存储过程的高效性也提高了导入流程的效率。

3.1.3 详解 cell 元素及子孙元素

每一个 cell 元素记录了报表中一个单元格的相关信息,描述这些信息的元素都作为 cell 的子元素:name 元素描述该单元格信息的名称;col 和 row 分别记录该单元格的列号和行号;pos 元素记录了该单元格信息对应在存储过程参数的位置。针对数据表中经常会出现的同类型数据在水平或者垂直方向延伸的特性,如图 3 所示,设计了 mode 元素和 length 元素分别记录数据延伸的模式和长度。导入程序会根据这两个值自动顺序读取一个或者多个单元格。

	A	B	C
1	样品编号	孔深(m)	
2		起	止
3	HZ17001	430.03	433.12
4	HZ17002	433.12	436.30
5	HZ17003	436.30	440.56
6	HZ17004	440.56	443.52
7	HZ17005	443.52	445.65

(a)垂直延伸

	A	B	C	D
1	记录孔深(m)	121.10	206.36	304.98
2	丈量孔深(m)	121.16	206.38	305.00
3	误差(m)	0.06	0.02	0.02

(b)水平延伸

图 3 报表中常见的数据组织方式

通过 Excel 编程接口读出的单元格信息都是字符型的,论文定义了原始类型返回给导入程序的数据类型,包括:字符型、整型和浮点型。返回值的类型通过 class 元素的属性值来描述。针对不同的返回类型, 分别定义了 T_Char、T_Float、T_Int 三种 class 的子元素。对某一个单元格,在确定的情况下 class 只能包含一种类型的子元素。

这三种元素都包含了 PreProcess 和 default 子元素。PreProcess 描述了从该单元格中提取信息之后要做的一些简单的字符串处理,可以包括 Replace(替换字符)、Delete(删除字符)、Trim(删除空格)操作。在定义元数据文件时,用户可以根据具体情况在 PreProcess 元素下面按照一定的顺序添加一个或者多个字符串操作,这些通用的字符串操作会被导入程序识别并执行。这种方式最大限度的给予了用户灵活性。default 元素定

义缺省值,当单元格为空时,用缺省值来替代空值。

此外,T_Char 元素还包含了 case 元素,用以描述该字符数据是否大小写敏感;T_Float 元素的子元素 ToAngle 描述是否将该浮点型按照 60 进制转换为角度值。图 4 是该元数据模型中主要元素之间的父子关系示意图。

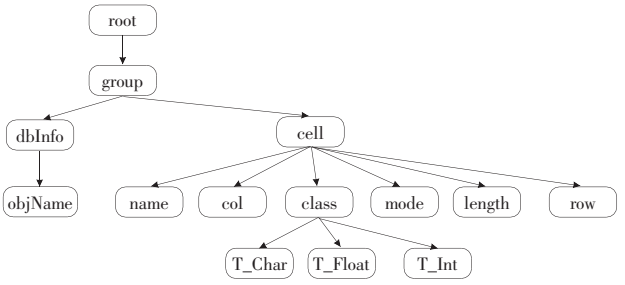

图 4　XML 元数据模型父子关系示意图

3.2　XML 和 Schema

上面描述的 XML 元数据模型层次较深、结构复杂,只有合乎上述规范的 XML 元数据文件才能被导入程序正确处理,因此,如何检验元数据文件的有效性显得尤为重要。针对该问题,使用 XML Schema 来规范 XML 元数据。

XML Schema 是 XML 的 Schema 语言, 用来描述 XML 文档的合法结构、内容和约束。XML Schema 由 XML 1.0 自描述,并且使用了命名空间,有丰富的内嵌数据类型和强大的数据结构定义功能,充分地改造并极大地扩展了 DTD(Documnet Type Definition)的能力。它正迅速替代 DTD 成为 XML 体系中正式的 Schema 语言, 与 XML 规范、Namespace 规范一起成为 XML 体系的坚实基础[3]。

在文中,利用 XML Schema 对元数据模型中元素的父子关系、名称、出现次数、出现顺序等进行了严格描述。除此之外,还描述了表 1 中所示信息。

导入程序读取 XML 元数据文件之后,会根据 XML Schema 的定义来校验元数据文件的有效性;只有有效的元数据文件才会被加入到其后的导入流程,否则系统会报错。这样,最大程度的保证了导入的正确性。

3.3　应用现状

采用上述元数据模型的导入系统采用 XML DOM Level2 Core 的类和 VC++实现。目前,该系统作为北京市城市地质信息管理与服务系统的一个核心模块,如图 5 所示,负责原始数据到数据库的自动化转储工作。

表 1　XML Schema 对元数据模型的定义

元素名称	约束内容
col	列号必须为包含[A..Z]的字母组合
row	行号必须为数字型
mode	枚举类型:CELL、VERTICAL、HORIZON 中任一
length	必须为数字型,可以为空
class	子元素为 switch 类型,子元素为 T_Char、T_Float、T_Int 中任一
pos	必须为数字类型,不能为空

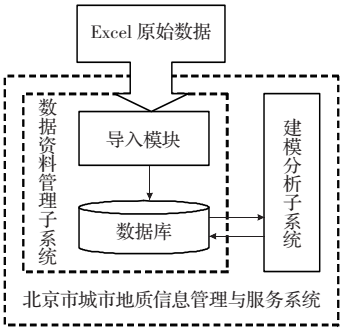

图 5　支持元数据的导入模块在系统中的应用

4　结束语

该元数据模型的可行性已经在实践中得到了很好的验证。原始数据对信息系统的重要性决定了该模型具有很强的通用性。目前,该模型对 Excel 数据表多工作簿情况下的处理多依赖人工判断,在今后的研究工作中有待于进一步完善。

参考文献:

[1] 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):93.

[2] 孙志东,潘懋,吴自兴,等.基于 XML 的地理信息元数据及空间数据安全[J].测绘通报,2007(9):61.

[3] 刘飞,黎建辉,阎保平.XML Schema 在科学数据库元数据互操作中的应用[J].计算机应用研究,2005(5):199.

(上接 107 页)

4　结语

实验证明,提出的基于 Bayes 推理的垃圾邮件特征选择评估函数,对于提高垃圾邮件过滤系统的性能与效率,进而提高整个过滤系统的实用性与可用性,效果非常显著,具有相当高的应用价值。

参考文献:

[1] Tan Pang-Ning,Stenbach M,Kumar V.数据挖掘导论[M].范明,范宏建,译.北京:人民邮电出版社,2006:13-50,137-150.

[2] Androutsopoulos I,Koutsias J,Chandrinos K V,et al.An evaluation of naive bayesian anti-spam filtering[C]//Proceedings of the Workshop on Machine Learning in the New Information Age,11th European Conference on Machine Learning.Barcelona,Spain,2000:9-17.

[3] 孙丽华,谢仲华,陈荣伶.信息论与纠错编码[M].北京:电子工业出版社,2005:14-36.

[4] Zorkadis V,Karras D A,Panayotou M.Efficient information theoretic strategies for classifier combination//feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering[J].Neural Networks,2005,18:799-807.

[5] Lai Chih-Chin.An empirical study of three machine learning methods for spam filtering[J].Knowledge-Based System,doi:10.1016/j.knosys.

[6] Stone T.Parameterization of naive bayes for spam filtering[R/OL].University of Colorado at Boulder,2003.http://trevorstone.org/school/spamfiltering.pdf.