基于 XML 映射模板实现不规则 Excel 数据的转换

武 彤,陆昱霖
(贵州大学 计算机科学与技术学院,贵州 贵阳 550025)

摘 要:不规则 Excel 数据与关系数据库数据的转换是很多企事业单位数据库应用系统在有效利用数据时经常碰到的棘手问题。文中研究的基于 XML 映射模板实现不规则 Excel 数据与关系数据库数据的转换方法,有效地解决了不规则 Excel 数据与关系数据库数据转换中存在的低效、繁琐、工作量大等问题。通过实际应用证明,所研究开发的转换系统可以批量转换不规则 Excel 数据,体现了高效性;并且可以为任何一个数据库应用系统处理不规则 Excel 数据提供帮助,体现了通用性及实用性。

关键词:可扩展标记语言;Excel 报表;不规则 Excel 数据;数据转换
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2015)07-0209-04
doi:10.3969/j.issn.1673-629X.2015.07.047

Realization of Conversion of Irregular Excel Data Based on XML Mapping Template

WU Tong,LU Yu-lin
(School of Computer Science and Information,Guizhou University,Guiyang 550025,China)

Abstract:The conversion between irregular Excel data and relational database data is a thorny problem to many enterprises and institutions in effective application of database. It conducts a research on the XML mapping template-based approach to implement the conversion between irregular Excel data and relational database data,which effectively improves work efficiency,simplifies working process,and is able to handle heavy work load. Through practical application of this conversion system to convert irregular data in batches,the efficiency has been proved. Furthermore,this conversion system can be widely applied to any database application system,which reflects the system's versatility and practicality.

Key words:XML;Excel statement;irregular Excel data;data conversion

0 引 言

随着信息技术的不断发展,企业之间的竞争不仅体现在产品的竞争上,更体现在对于关键信息的管理和利用上。OA 系统的运用使得企事业单位的管理发生了前所未有的改变,办公效率得到了空前提高。其中 Excel 报表作为企事业单位上传和下达的重要信息载体,随着信息化建设的不断推进,在工作中得到了广泛的应用[1]。

Excel 报表数据的管理已经成为企事业单位办公人员工作中不可或缺的部分,处理 Excel 报表数据的方便性和灵活性成为评价 OA 系统设计成功与否的重要标准。同时很多数据库应用系统通过导入 Excel 数据进行分析和挖掘,找出隐藏在数据中的知识和规律以支持决策。根据大量统计资料表明,30% 左右的规则 Excel 数据有效地存储在各种类型的关系数据库管理系统中,但还有 70% 左右的不规则 Excel 数据分散在整个业务过程及外部环境中。

不规则 Excel 数据作为企事业单位做决策的重要依据,怎样有效地管理好这些不规则 Excel 数据,并挖掘出这些数据的内在联系和重点知识是目前急需解决的问题[2]。

文中通过将 XML 文档作为实现不规则 Excel 数据与关系数据库之间转换的桥梁,研究基于 XML 映射模板实现不规则 Excel 数据与关系数据的转换,从而实现有效利用 Excel 数据为企事业单位进行科学决策服务。

收稿日期:2014-09-01 **修回日期**:2014-12-03 **网络出版时间**:2015-06-23
基金项目:贵州省科技攻关项目(黔科合 GY 字【2010】3061)
作者简介:武 彤(1964-),女,教授,硕士,CCF 会员,研究方向为数据仓库技术、OLAP、数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1046.033.html

1 XML 映射模板的设计

XML(eXtensible Markup Language,可扩展标记语言)是由 WorldWide Web Consortium(W3C)的 XML 工作组定义的。这个工作组是这样描述该语言的:XML 是 SGML(Standard Generalized Markup Language,标准通用标记语言)的子集,其目标是允许普通的 SGML 在 Web 上以目前 HTML(Hyper Text Markup Language)的方式被接收、服务和处理。XML 被设计成易于实现,且可在 SGML 和 HTML 之间相互操作[3]。XML 之所以成为事实上的网络数据表示标准,是因为它具有其他表示方式所不具备的特点,主要有[4]:

(1)结构化。

XML 是结构化的语言规范,遵循严格的语法要求。XML 文档附有一个 DTD 规范,DTD 定义了 XML 文件的语法和数据结构,将 XML 的内容与结构分离。

(2)自描述。

XML 具有自描述信息。XML 文档一般包含一个文档类型声明,以便能让使用者根据需要定义属性名和标记,也可以通过描述语法描述 XML 文件的结构。XML 对数据的描述方式是独立于应用系统的,所以可以实现数据的重用。可以把 XML 文档看作是文档的数据库化或者数据库文档化。同时 XML 的自描述性也提高了 Web 的检索功能。

(3)可扩展。

XML 继承了 SGML 易扩展的特性。XML 允许用户自行定义,可以使用自己的标记也可以使用他人共享的标记。HTML 只能使用预先定义好了的固定的标记集合来描述 Web 的所有数据元素,所以 HTML 的扩展性不好,不能在不破坏 HTML 标准的情况下增加新的标签。然而现在的 Web 页面需要表达的数据内容日渐复杂丰富,根据 Web 页面发展的需要,新的页面标签需要具有良好的可扩展性。XML 由 DTD 定义标签,允许用户根据实际需要,自己创建标记集和文档结构以描述 Web 页面的任何数据元素。各种应用程序可以根据这些标记理解文档中的数据。XML 越来越多地应用在了 Web 应用中。

1.1 关系模式与 XML 模式

19 世纪 70 年代由美国著名的 IBM 公司的研究员 E. F. Codd 首次提出了数据库系统的关系模型,该理论开创了数据库关系方法和关系数据理论研究的先河。自此以后以关系模型为基础的关系数据库系统得到了广泛的应用。关系模型以严格的数学理论为基础。关系模型是规范化的模型,即要求关系必须满足一定的规范条件,这些规范条件中最基本的一条是:关系模型中要求关系中的任何一个分量都必须是不可再分的数据项,也就是要求不允许表内还嵌套表[5]。

XML 的组织单位是 XML 文档,XML 文档是由根元素和很多子元素构成的。每个子元素包含许多内容甚至还能嵌套其他子元素。元素用标记来标识和界定,每个元素由一个开始标记和一个结束标记配套进行标识。从数据管理的角度来分析 XML 文档,可以认为 XML 文档是具有层次结构的半结构化数据。而关系模型是为存储和管理结构化数据设计的,关系模式存取数据的形式是扁平的二维表[6]。

XML 模式与关系模式之间存在固有的不匹配性。如果要把 XML 数据存储到关系数据库中,首先必须为 XML 文档创建一个关系模式,这样就把问题转化为两个异构模式之间的模式映射问题。

1.2 XML—关系模式的映射方法

XML 模式到关系模式的映射方法可以根据关系模式的生成方法分为以下五类:

(1)基于边或属性产生关系模式:该方法通过把 XML 实例模型化成一个有向图,然后再利用基于边或属性的映射方法产生关系模式[7]。

(2)从 XML 文档实例中提取关系模式:首先根据 XML 文档实例产生 XML 模式,然后把 XML 模式映射为关系模式。

(3)从文档模式定义导出关系模式:对 XML 文档模式定义中的语法进行分析后映射为关系模式[8]。XML 文档模式定义包括 DTD、XML Schema 等。

(4)从文档模式定义及 XML 实例数据导出关系模式[9]:该方法综合了上述(2)和(3)两种方法,不仅利用了 XML 的模式定义文档,而且利用了 XML 文档实例信息,所以对于产生好的关系模式提供了充分的条件。

(5)系统缺省或用户定制的模式映射方法:这种方法被很多商业 RDBMS 所采纳。系统首先定义一个缺省的映射方案,然后让客户根据需求自己修改这个已定义的映射方案或者客户用系统提供的工具自行定义映射[10]。

1.3 XML 映射模板设计方法

由上面分析可知,XML 映射模板是不规则 Excel 数据与关系型数据库数据转换的桥梁[11]。所以 XML 映射模板的设计是文中所研究系统成功与否的关键一步。文中研究开发的系统在实现 XML 到关系数据库的映射时,采用的方法是从 XML 文档出发将原文档直接映射为关系模式,而不再考虑 XML 的其他模式信息。采用这种方法因为其简单易用,编程实现较快,且使用范围更广,对于很多不懂 XML 的开发人员也能进行良好的维护。

由于 XML 映射模板的设计目的是辅助实现不规则 Excel 数据与关系型数据库数据的转换,所以设计

XML 必须要保存不规则 Excel 数据信息以及结构信息还有关系型数据库信息,因此需要利用 XML 的自描述性,通过自定义合适的 XML 节点标签,将这些信息存储起来,然后形成不规则 Excel 数据与关系数据库之间的映射关系,再利用该映射关系实现不规则 Excel 数据与关系数据库之间的转换。通过分析一般不规则 Excel 数据与关系数据库的结构特点,发现 XML 映射模板只需要记录各个数据的坐标位置、数据之间的对应关系以及关系表的信息等关键内容就能辅助实现不规则 Excel 数据与关系数据库数据的转换。以下代码就是不规则 Excel 数据转换系统的 XML 映射模板,通过该 XML 映射模板实现不规则 Excel 数据与关系数据库之间的转换[12]。

```
<DataConfig>
  <TableName>TableName</TableName>----------该节点记录数据需要存储的关系表名
  <FieldDesc>----------该节点记录处理数据的信息
   <DefinitionDesc>----------该节点记录的数据在关系数据库的映射信息
    <RowIndex>RowIndex</RowIndex>-----该节点记录该字段数据所在 EXCEL 表的行坐标
    <ColumnIndex>ColumnIndex</ColumnIndex>该节点记录该字段数据所在EXCEL表的列
    <FieldName>FieldName</FieldName>--------该节点记录该数据在关系数据库表的字段名
    <DataType>DataType</DataType>--该节点记录该数据在关系数据库表的数据类型
   </DefinitionDesc>
    <ValueDesc>----------该节点记录的是上面关系数据库字段值的坐标
      <RowIndex>RowIndex</RowIndex>----------该节点记录该字段值的行坐标
      <ColumnIndex>ColumnIndex</ColumnIndex>------该节点记录该字段值的列坐标
    </ValueDesc>
  </FieldDesc>
</DataConfig>
```

2　不规则 Excel 数据与关系数据转换的实现

文中研究利用 XML 映射模板实现了不规则 Excel 数据与关系数据库的转换,最终构建了一个通用的转换系统。具体实现功能如下所述。

2.1　XML 映射模板的生成

XML 映射模板的生成功能是指用户通过前台界面自定义生成某类型不规则 Excel 表的 XML 映射模板,目的是实现同类型不规则 Excel 报表到关系数据库的批量快速转换,这是实现不规则 Excel 数据与关系数据库转换的关键。用户进入系统主页面后,选择要处理的 Excel 文件,然后进入映射模板生成功能页面。用户根据自己需求,通过交互方式操作,进行 XML 映射模板的生成,操作完成后,应用程序将会生成相应的 XML 映射文件,并将生成的 XML 映射文件信息存入数据库,以备下次转换同种类型的不规则 Excel 数据时使用。

功能实现流程如图 1 所示。

2.2　关系数据库表的生成

关系数据库表生成功能是实现不规则 Excel 数据转换成关系数据库数据的前提,必须要生成相应的数据库表,才能将需要转换的不规则 Excel 数据存储起来,以便于后期的查询和统计分析。现阶段大部分处理不规则 Excel 报表的方法往往需要手动先建立关系数据库表,再将不规则 Excel 报表处理成规则 Excel 报表,最后才能将相应的规则 Excel 数据信息导入到关系数据库[13]。由于关系数据库表的创建和不规则 Excel 报表的导入是不同步的,所以处理难度会加大而且容易出现意想不到的错误,并且会增加工作量。文中研究的不规则 Excel 数据与关系数据库数据转换系统,通过同步生成 XML 映射模板与数据库关系表,不仅提高了工作效率,而且避免了不规则 Excel 数据导入到关系数据库表时出现其他错误。

功能实现流程如图 2 所示。

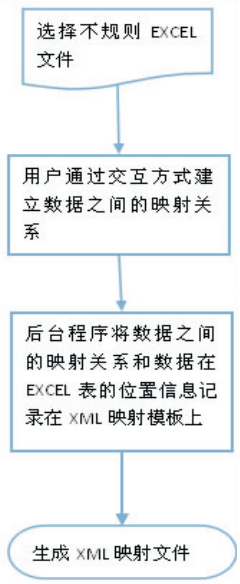

图 1　XML 映射模板的生成流程图

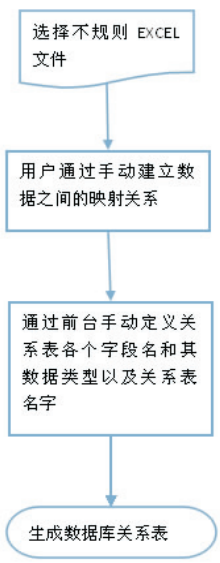

图 2　数据库关系表

2.3　不规则 Excel 数据导入关系数据库

不规则 Excel 数据导入到关系数据库功能是在对

应的 XML 映射模板和数据库关系表都已经存在的前提下,利用此功能便捷地将不规则 Excel 数据直接导入到关系型数据库而不需要重复进行 XML 映射模板的生成工作。此功能改变了现在大多数处理不规则 Excel 报表时,需要重复地进行不规则 Excel 报表手动加工成规则 Excel 报表再进行数据导入存储的现状。本系统在处理以往已经处理过,并且已经存在相同格式的 XML 映射模板的不规则 Excel 报表时,只需要直接选择其对应的 XML 映射模板就能进行不规则 Excel 报表数据的导入[14]。

功能实现流程如图 3 所示。

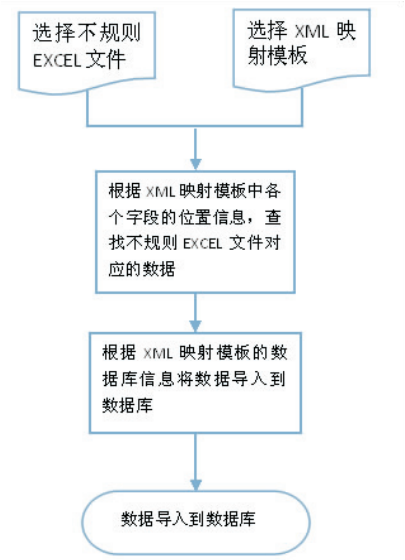

图 3 不规则 Excel 数据导入到关系数据库

3 不规则 Excel 数据转换系统的特点

文中研究开发的不规则 Excel 数据与关系数据转换系统与传统处理不规则 Excel 数据与关系数据的转换方法有所不同,体现了如下特点:

(1)高效性。

文中研究实现的系统提出了以 XML 技术作为不规则 Excel 数据与关系数据库数据转换的桥梁。系统通过为不规则 Excel 报表生成一个 XML 映射文件,然后再利用该 XML 映射文件将不规则 Excel 数据转换成关系数据库数据。这种处理方法,实现相对简单,开发成本较低,并且能够满足大部分单位的实际需要。由于第一次转换不规则 Excel 报表时,保留了 XML 映射模板,所以当再次处理相同格式的不规则 Excel 数据时就能直接利用已生成的 XML 映射模板将不规则 Excel 数据直接导入到关系数据库,能实现批量导入不规则 Excel 数据,大大提高了对不规则 Excel 数据与关系数据库的转换效率,从而实现了不规则 Excel 数据的高效管理。

(2)通用性。

文中研究开发的系统并不针对某一家单位或某一个系统。该系统是为解决不规则 Excel 数据与关系数据库数据转换这一实际问题而开发设计的通用软件,所以该系统的设计目标之一就是可以被任何单位或者任何数据库应用系统直接使用,解决利用 Excel 数据中碰到的实际问题,所以通用性也是该系统的重要特点。

4 结束语

文中研究的不规则 Excel 数据与关系数据库转换系统,主要利用 XML 映射模板实现了不规则 Excel 数据与关系数据的转换。该系统对于批量导入相同格式的 Excel 数据,与传统的处理方法相比,体现出较高的转换效率。系统的通用性体现在可以为任何一个数据库应用系统处理不规则 Excel 数据时提供帮助。经过企业的实际应用,证明了该系统具有一定的实用性及推广使用价值。

参考文献:

[1] 刁瑜平. 人力资源管理系统的研究和实现[D]. 广州:广东工业大学,2007.
[2] 尹艳梅. 我国企事业单位信息化建设的制约因素与对策[J]. 企业家天地,2011(4):12-13.
[3] 耿祥义,张跃平. XML 基础教程[M]. 北京:清华大学出版社,2012.
[4] 裴 松. 半结构化数据处理方法在生产线控制系统中的应用[D]. 贵阳:贵州大学,2014.
[5] 王 珊,萨师煊. 数据库系统概论[M]. 北京:高等教育出版社,2009.
[6] 杨小兵. 基于关系数据库的 XML 电子病历系统研究[D]. 武汉:华中科技大学,2009.
[7] Yoshikawa M, Shimura T, Uemura S. Xrel: a path-based approach to storage and retrieval of XML documents using relational database[J]. ACM TOIT, 2001, 1(1): 110-141.
[8] 樊艳芬. XML 文档关系存储技术之模式映射方法的研究[D]. 南昌:江西师范大学,2007.
[9] 乔璞雪. 关系数据库中 XML 数据挖掘技术研究与实现[D]. 天津:南开大学,2010.
[10] Chu W W. CPI: constraint-preserving inlining algorithm for mapping XML to relational schema[J]. Data and Knowledge Engineering, 2001, 39(1): 3-25.
[11] 宫 勋. 关系型与 XML 数据库格式转换方法及应用[D]. 大连:大连理工大学, 2008.
[12] 宋 磊. 电子政务平台下数据转换技术的研究[D]. 北京:北方工业大学,2007.
[13] 文 龙. 基于 XML 的非结构化数据管理研究及应用[D]. 长沙:湖南大学,2009.
[14] 陆昱霖. 不规则 EXCEL 数据与关系数据库转换方法的研究及实现[D]. 贵阳:贵州大学,2014.

基于XML映射模板实现不规则Excel数据的转换

作者： 武彤， 陆昱霖， WU Tong， LU Yu-lin
作者单位： 贵州大学 计算机科学与技术学院,贵州 贵阳,550025
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(7)