第 34 卷 第 23 期 计 算 机 工 程 2008 年 12 月
Vol.34 No.23 Computer Engineering December 2008
·软件技术与数据库· 文章编号：1000—3428(2008)23—0065—03 文献标识码：A 中图分类号：TP311.5

Excel 报表数据的自动分离引擎

白仲贵，张毅坤，杨凯峰，邓晶晶，王 凯
(西安理工大学计算机科学与工程学院，西安 710048)

摘 要：针对当前 Excel 报表数据自动采集的局限性，提出一种 Excel 报表数据自动分离的方法，并以此为基础进行引擎的研究。该引擎采用两级映射(模板样式到模板结构树和模板结构树到 XML 架构)，根据 Excel 模板自动生成 XML 架构和映射信息。借助 Excel 数据分离机制将 Excel 报表数据自动分离成与模板样式相对应的 XML 数据文件，使 Excel 报表的数据采集更加容易，更加有利于系统扩展与集成。

关键词：引擎；XML 架构；模板结构树；数据自动分离

Automatic Separation Engine of Excel Report Data

BAI Zhong-gui, ZHANG Yi-kun, YANG Kai-feng, DENG Jing-jing, WANG Kai
(School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048)

【Abstract】 With the limited functionality of data automatic collection in Excel, this paper suggests a way to automatically separate data in Excel reports and discusses engine. This engine uses two layer mapping (template maps to template structure tree, and template structure tree maps to XML schema). XML schema and the mapping path of each cell are generated automatically from Excel template. The related XML document which is in accord with template is automatically separated from Excel report data by the Excel data separation mechanism and it will be much easier to collect data from Excel report and be more advantageous to system's expansion and integration.

【Key words】 engine; XML schema; template structure tree; data automatic separation

1 概述

由于Microsoft Excel报表的普遍性，大多数企业信息系统如OA系统、ERP系统和MIS系统等都离不开与Excel报表进行交互，实现数据采集。目前对Excel报表数据采集方法大体分为 2 种[1]：(1)直接通过Office PIA访问Excel，这种方法最大的局限性在于系统要针对每个工作表进行编程，系统开发工作量大，而且不易扩展和维护。(2)将Excel工作表视为数据表，而将工作薄视为数据库，通过标准的数据访问策略来访问Excel数据。这种方法最大的局限性在于要求Excel工作表为二维表。

为了解决上述 Excel 报表数据采集的局限性，本文提出了一种借助于 Microsoft Excel(以下简称 MS Excel) XML 数据分离机制，实现 Excel 报表数据自动分离的方法，并以此为基础进行 Excel 报表数据自动分离引擎的研究。该引擎根据 Excel 报表模板自动生成 XML 架构和映射信息 ;实现了 Excel 报表数据到 XML 数据的自动转换 ;克服了 Excel 数据分离需要人工干预，并要求操作人员掌握 XML 架构编写方法的弊端，使企业信息系统对 Excel 报表数据实现自动采集。

2 MS Excel XML 数据分离机制简介

2.1 MS Excel 的 XML 映射

在 Excel 2003 中新增了 XML 映射功能，可以将 XML 架构映射到 Excel 工作簿。映射时，可以从任何符合映射架构的 XML 源中导入导出数据。因此，利用 XML 映射功能，将 XML 架构与工作簿相关联，这样能够更为简单和可靠地在 Excel 中导入和导出数据。

2.2 MS Excel的XML数据分离方法[2]

首先，根据实际需要编写自定义 XML 架构；其次，将已经编写好的 XML 架构添加到已打开的工作簿中；然后，使用“ XML 源 ”将单元格映射到架构元素，就可以向映射的单元格中无缝导入或从中导出 XML 数据。

3 Excel 报表数据自动分离引擎的体系结构

3.1 Excel 报表数据自动分离引擎的分离方法

根据 MS Excel 提供的数据分离机制，可以看出，要实现数据自动分离，必须解决 XML 架构和映射信息的自动生成问题。

经过认真分析发现，要解决以上问题必须从模板样式着手，根据模板样式自动生成 XML 自定义架构和映射信息；同时借助 Excel 数据分离机制，实现 Excel 报表到 XML 数据的自动分离。

3.2 Excel 报表数据自动分离引擎体系结构

Excel报表数据自动分离引擎主要用来实现将Excel报表数据自动分离成XML数据和将XML数据与Excel模板动态组合从而实现数据动态查询。要实现报表数据自动分离和动态查询首先要进行模板注册。模板注册主要为了生成 3 方面信息：模板的基本信息，数据分离所需要的自定义XML架构和将单元格与架构元素对应的映射信息，这 3 方面信息是实现数据自动分离和动态查询的基础。数据自动分离引擎的总体体系结构如图 1 所示。

基金项目：陕西省自然科学基金资助项目(2005F07)
作者简介：白仲贵(1980－)，男，硕士研究生，主研方向：软件工程，软件自动化；张毅坤，教授；杨凯峰，讲师；邓晶晶、王 凯，硕士研究生
收稿日期：2008-04-06 **E-mail：**bzg_yxy@163.com

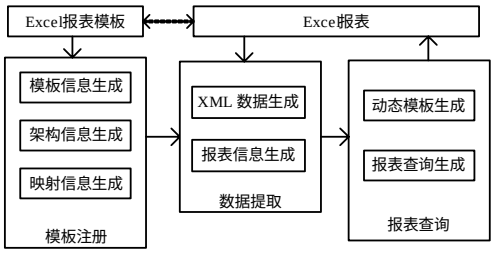

图 1　Excel 报表数据自动分离引擎体系结构

4　关键技术实现

4.1　架构信息和映射信息自动生成

这是 Excel 报表数据自动分离引擎能够实现数据自动分离的关键所在。

4.1.1　架构信息自动生成算法的核心思想

首先将 Excel 模板样式用树型结构表示，而把每一个数据单元格作为树的叶子节点；其次根据模板样式结构树生成自定义 XML 架构，即实现两级映射：一是从模板样式到模板结构树的映射；二是模板结构树到 XML 架构的映射。

例如：煤矿掘进进尺表模板样式如图 2 所示，其中字母 A~L 不是样式的组成部分，而是作为数据单元格标识，即数据单元格的名称。

3		计划		实际完成	%		提前完成计划天数
4		考核	作业		考核	作业	
5	全矿	A	B	C	D	E	F
6	其中:开拓	G	H	I	J	K	L

图 2　煤矿掘进进尺 Excel 模板样式

实现从模板样式到模板结构树第一级映射所产生的模板结构树如图 3 所示。

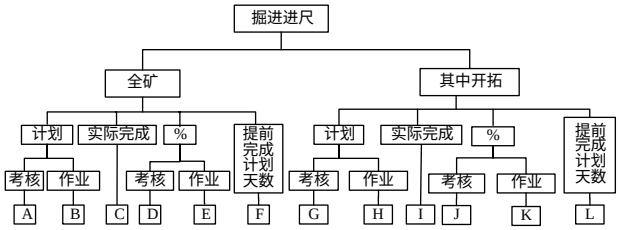

图 3　掘进进尺模板结构树

实现模板结构树到 XML 架构第 2 级映射所生成的 XML 架构信息如下所示：

<xs:schema> <xs:element name="掘进进尺"><xs:complexType>
<xs:sequence><xs:element name="全矿"><xs:complexType>
<xs:sequence><xs:element name="计划"><xs:complexType>
<xs:sequence><xs:element name="考核" type="xs:string" />
<xs:element name="作业" type="xs:string" /></xs:sequence>
</xs:complexType></xs:element>… </xs:complexType>
</xs:element><xs:element name="其中开拓"><!—“其中开拓”的架构与“全矿”完全一致,这儿略--></xs:element></xs:sequence>
</xs:complexType> </xs:element></xs:schema>

4.1.2　实现 2 级映射的算法设计

为实现从模板样式到模板结构树第一级映射的算法，首先要获取模板的基本信息，模板基本信息包括：

(1)模板类型：通过对大多数 Excel 报表分析可知，Excel 报表数据区域大概分为 3 类，即顺序表、交叉表和分类汇总表。顺序表的特点是上方为坐标项，下方为数据；交叉表的特点是中间部分为数据，上方和左侧为其坐标项，其中上方坐标项通常称为水平坐标项，左侧坐标项通常称为垂直坐标项；分类汇总表的特点是一种特殊的交叉表，每一个分类项有一个合计项。

(2)辅助坐标类型：在一些报表系统中，为了简化复杂报表(多维报表)的生成，报表设计者使用了辅助项将复杂报表变为二维报表。辅助坐标项分为 4 类：行坐标，列坐标，两者皆有或两者皆无。

(3)模板数据区域信息：模板数据区域以封闭区域为标志，即处于封闭区域的单元格为数据区域。模板数据区域信息包括：数据区域起始单元格的位置，即在工作表中的起始位置(行坐标号和列坐标号)；数据区域坐标项的行列数，即水平坐标项的列数和垂直坐标项的行数以及数据区域的行数和列数。

其次，根据模板基本信息使用递归算法对整个数据区域的坐标项进行深度优先遍历，动态生成模板结构树，其主要思想如下：

(1)如果模板类型为交叉表。1)将数据区域起始单元格作为模板结构树的根，当该起始单元格为空，则根节点名称可以自拟，如项目或模板名称。而且起始单元格正左下方单元格作为该模板结构树的第 1 级子节点。2)当水平坐标项列数为 2，则第 1 级子节点单元格(合并单元格)正右方单元格为该单元格的子单元格，即为该树的第 2 级子节点。当水平坐标项列数为 3，依此类推。3)当水平坐标项遍历到最右单元格时，开始遍历垂直坐标项，而把垂直坐标项作为水平坐标项最右方各坐标单元格的子孙单元格，进行遍历。对垂直坐标项进行遍历类似于对水平坐标项进行遍历，唯一区别在于水平坐标项向右优先遍历，而垂直坐标项向下优先遍历。此类模板结构树动态生成递归算法流程如图 4 所示。

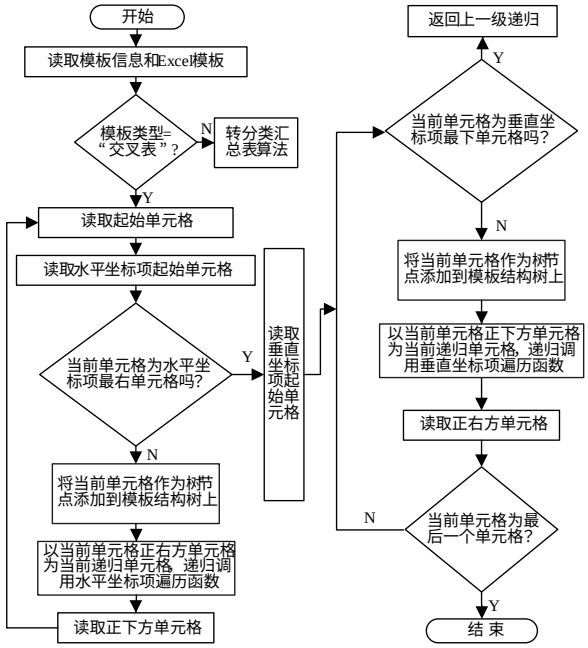

图 4　交叉表模板结构树动态生成递归算法流程

(2)如果模板类型为分类汇总表。对分类汇总表进行遍历类似于对交叉表的遍历，只不过还要对分类汇总项进行遍历，具体遍历算法略。

(3)如果模板类型为顺序表。直接开始遍历垂直坐标项。此模板结构树的根节点名称可以自拟，如项目或模板名称。

实现从模板结构树到自定义 XML 架构的第 2 级映射的算法：

(1)如果是基于.NET Framework 开发程序，可使用 XML 架构定义工具(Xsd.exe)直接将模板架构树的 XML 表示形式转换为相对应的 XML 架构。

(2)如果是基于其他平台开发程序，则可通过编程的方式根据模板架构树的 XML 表示转换为它所对应的 XML 架构，具体算法略。

4.1.3　映射信息自动生成

在实现从模板样式到模板结构树第 1 级映射过程中，通过向递归算法中加入一些记录单元格映射信息的参数，自动生成数据单元格映射信息，如下所示：

<CellMapInfo> <CellMap><row>5</row><column>2</column> <xpath>/项目/全矿/计划/考核</xpath> <repeat>false</repeat> </CellMap> <CellMap><row>5</row><column>3</column> <xpath>/项目/全矿/计划/作业</xpath><repeat>false</repeat> </CellMap>…</CellMapInfo>

4.2　数据自动分离[3]

引擎借助 Excel XML 数据分离机制实现数据自动分离。具体方法是：通过调用 Excel COM 组件中的部分对象来实现数据自动分离。

4.2.1　加载架构

在 Excel 中，XmlMap 对象代表一个或多个架构及其到电子表格的映射。以下代码将新架构添加到 Workbook 对象：

Application.Workbooks(1).XmlMaps.Add("c:\schemas\schema.xsd")

其中，"c:\schemas\schema.xsd"架构文件通过引擎模板注册组件自动生成。

4.2.2　建立映射

向工作薄中添加架构以后，接下来需要将单元格或范围映射到 XML 架构中的元素。例如：

ActiveSheet.Range("B6").XPath.SetValueActiveWorkbook.XmlMaps(1), "/区队名称/原煤产量/单位"

引擎根据模板注册组件所产生的数据单元格映射信息自动建立映射。

4.2.3　导出数据

加载架构和建立映射后，可以使用 Workbook 对象的 Export 方法，为任何加载的架构创建文档。例如：

ActiveWorkbook.XmlMaps(1).Export _"c:\data\data.XML"

其中，文档"c:\data\data.XML"为数据自动分离引擎最终所要分离出的结果。

5　应用举例

通过利用引擎模板注册组件将图 2 煤矿掘进进尺模板进行注册，将产生如前所述的 XML 架构信息和映射信息。

模板注册以后，利用引擎数据提取组件将图 5 所示的 Excel 报表数据进行分离，所分离出来的 XML 数据如下所示：

<掘进进尺><全矿><计划><考核>100</考核><作业>100</作业> </计划><实际完成>120</实际完成><百分比> <考核>120</考核> <作业>120</作业></百分比><提前完成计划天数>5</提前完成计划天数></全矿>……</掘进进尺>

3		计 划		实际完成	%		提前完成计划天数
4		考核	作业		考核	作业	
5	全矿	100	100	120	120	120	5
6	其中:开拓	100	100	120	120	120	5

图 5　煤矿掘进进尺 Excel 报表

另外，基于此引擎笔者成功地实现了某煤矿集团公司基于 MS Excel 的通用报表管理系统。

6　结束语

本文借助 MS Excel XML 数据分离机制，提出一种 Excel 报表数据自动分离的方法，并以此为基础进行 Excel 报表数据自动分离引擎的研究。此分离引擎将 Excel 报表数据自动分离成与模板样式相对应的 XML 数据文件，迎合了当今软件开发技术的潮流，易于系统扩展与集成。对一些不规范的报表进行数据自动分离还有待进一步研究。

参考文献

[1] 郑宇军, 朱连军. 新一代.NET Office 开发指南——Excel 篇[M]. 北京: 清华大学出版社, 2006.

[2] Walkenbach J. 中文版 Excel2003 宝典[M]. 陈　缅, 裕　鹏, 译. 北京: 电子工业出版社, 2004-05.

[3] Vogel P. 使用 Excel 2003 对象模型添加 XML 数据集成[EB/OL]. (2004-08-01). http://www.microsoft.com/china/msdn/library/office/office/odcxlExcel2003XMLIntro.mspx?mfr=true.

~~~~~~

（上接第 64 页）

本文结合实际的应用，深入研究并实现了基于中文的常问问答系统的构建极其在实际中的应用。工作包括：充分利用了知网的知识资源，实现了基于知网的相似度计算方法，并分析了该算法的利弊；结合中文语义丰富的特点，引入了词义消歧算法，实现了基于知网的词义消歧算法，并进行了改进。实验结果表明，改进后的消歧算法取得了较好的效果。将该算法引入到基于知网的词义消歧算法，可以提高系统的速度和精度；实现了改进后的相似度算法，并将其应用到实际的 FAQ 系统中，实现了真正的语义理解。实验证明：引入了词义消歧后，FAQ 系统的性能可以达到实际应用的要求。

未来的工作，主要集中在词义消歧算法的改进及消歧语料的完善方面。此外，HowNet 辞典词条的收录等方面也是影响该系统性能的一个重要因素。因此，针对不同目的的 FAQ 系统，辞典文件的选取也显得尤为重要。如引言所述，汉语是很复杂的，中文信息处理才刚刚起步，仍面临着很大的挑战。因而，在中文自然语言处理领域中，很多细节问题还有待进一步研究。

参考文献

[1] Chatterjee N. A Statistical Approach for Similarity Measurement Between Sentences for EBMT[C]//Proc. of Symposium on Translation Support Systems. [S. l.]: IEEE Press, 1999.

[2] 车万翔, 刘　挺, 秦　兵. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯, 2004, 14(7): 15-20.

[3] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林[M]. 2 版. 上海: 上海辞书出版社, 1996.

[4] 丁江伟. 基于依存分析的全文词义消歧研究[D]. 哈尔滨: 哈尔滨工业大学, 2002.

[5] 杨尔弘, 张国清, 张永奎. 基于义原同现频率的汉语词义排歧方法[J]. 计算机研究与发展, 2001, 38(7): 833-838.

[6] 崔　桓, 蔡东风, 苗雪雷. 基于网络的中文问答系统及信息抽取算法的研究[J]. 中文信息学报, 2004, 18(3): 24-31.
~~~~~~