网络环境下基于 XML 的复杂格式数据传输的方法①

戚婷婷　李小林 (中国科学技术大学 管理科学与工程系 安徽 合肥 230026)

摘　要： 提出了一种在网络环境下基于 XML 实现客户端与服务器之间复杂格式数据传输的方法。该方法借助 Microsoft Word 2003 (及以上版本)对 XML 语言的支持，利用 XML 语言文本格式的特点将复杂格式的数据信息转换为文本信息，经过合理分析和组织后对其进行保存和重现，使用户能够在 B/S 模式下实现对复杂格式数据的处理。采用该方法的在线文档生成系统已由程序实现。提供了该方法的实现思路及关键部分的处理。

关键词： 数据传输；复杂格式；可扩展性标记语言(XML)

A Method to Transmit Complex Formatting Data Based on XML in Network Environment

QI Ting-Ting, LI Xiao-Lin (Department of Management Science and Engineering, University of Science and Technology of China, Hefei 230026, China)

Abstract: This paper proposes a new method to transmit complex formatting data for Client/Server mode based on XML. Complex formatting data are transformed into text data by applying characteristics of XML according to this approach. These text data can be reorganized and represented to fulfill the requirements of data processing. The online documents generation system with this approach has been implemented. The idea and critical parts of this approach are addressed.

Keywords: data transmission; complex formatting; XML

1　引言

通常网页上的文本控件可以使用户轻松实现从客户端向服务器提交文本数据的功能，并给文本添加丰富的样式。但有时用户所希望提交的信息并不局限于简单的图文数据，而是除图文之外的形状、表格或公式等复杂格式的数据。由于在日常办公中大多数用户已习惯使用 Microsoft Word 程序来处理和传递信息，并且 Word 程序编辑功能强大，而目前的在线文本处理软件功能与 Word 程序相比缺乏多样性，故本文提出一种在客户端实现功能同于 Word 程序中对复杂格式数据进行处理及传输的方法。

2　相关技术介绍

2.1 部分在线文档编辑器的比较

目前，用户主要使用在线文本编辑器进行网页数据的提交，这种方式简便有效，能够完成简单格式的图文数据的处理及传输，但无法对复杂格式的图文数据进行操作。所谓复杂格式的图文数据操作是指 Word 程序可能实现的一切功能，如对文本数据的控制及对图片、形状和公式等的编辑。由于本文所讨论的问题主要针对办公自动化的情况，故针对以下两类软件进行比较。

一类是通过 ActiveX 控件调用 Word 程序进行相

① 收稿时间:2009-06-23

应的文档编辑，如 NTKO OFFICE 文档控件和 DSOFramer(Developer Support Office Framer Control)控件。其中，NTKO OFFICE 文档控件能够在浏览器窗口中直接调用 word,excel,wps 等 office 文档并保存到 Web 服务器。由于该控件能够直接调用 Word 程序，因而实现了 Word 软件的所有功能，最大限度地满足了用户对复杂格式网络文档编辑的要求，是一种将办公自动化及文档管理与 Microsoft office 等软件集成起来的解决方案,是办公自动化系统的最佳控件之一。但该控件价格较高，故在使用上有不便之处。类似软件还有金格 office 控件、办公之星等。对于 DSOFramer 控件，它作为 ActiveX 文档容器，用于承载用户窗体或 Web 网页中的 office 文档(如 Microsoft 系列的 word,excel, powerpoint, project 及 visio 等),具有操作灵活和轻量级的特点,为开发人员在自己的解决方案中使用 office 文档提供了新的技术手段[1]。由于该控件是开源的，国内的技术人员在源码的基础上进行了改进，增强了该控件的功能，因此该控件使用的局限性较小，功能性较强。

另一类是基于网页的在线文本编辑器，如 FCKeditor,Google 的在线文档等。FCKeditor 是一款功能强大的开源在线文本编辑器，能够使用户基于 Web 实现类似于桌面文本编辑器 Microsoft Word 的许多文档编辑功能。由于它不需在客户端进行安装，同时兼容大多数主流浏览器，具有轻量级的特点，易于用户通过网页对文本及图片进行控制。FCKeditor 完全基于 Javascript 来实现，使用 HTML 语言对文本进行控制，使用方便、加载快捷。

2.2 XML 与 Microsoft office 介绍

可扩展性标记语言 (eXtensible Markup Language，缩写为 XML)是由 WWW 联合会(W3C)于 1998 年 2 月发布的一种标准，是针对网络应用的一项新技术。XML 是一种优化的标准通用标记语言(SGML)，它实现了内容和形式的分离，具有良好的扩展性、跨平台移植性和自描述等特点。 XML 文档是纯文本，可以使用文本编辑器和可视化开发环境的任何工具进行创建和编辑。XML 对格式的定义灵活，具有很好的层次结构，标记可以嵌套，其优势主要体现在互联网的数据交换[2]。典型的 XML 文档结构如例程 1 所示。

例程 1 典型的 XML 文档结构

```
<?XML version="1.0" standalone="yes"?>
<Contents>
    <Chapter>
        <ChapterName>a</ChapterName>
        <SecteName>b</SecteName>
    </Chapter>
  …
</Contents>
```

由于 XML 的自定义性及可扩展性，使其可以做到数据显示与内容的分离。另外，XML 还可以在不同的系统或应用程序之间进行数据交换，从而解决了数据间的接口问题，使得 Word 程序与其他应用程序之间的数据交换有了统一的接口[3]。

2.3 ASP.NET 及 DCOM 接口

ASP.NET 是 ASP(Active Server Pages)的后继版本，ASP.NET 及其前期版本都是构建新一代动态网站和基于网络(特别是 Internet)的分布式应用的技术。ASP.NET 为网站设计人员和网络程序员提供了更简单快捷的开发方法。

分布式组件对象模型(DCOM)是一系列微软的概念和程序接口，使得客户端程序对象能够请求来自网络中另一台计算机上的服务器程序对象。DCOM 基于组件对象模型(COM)，提供了一套允许同一台计算机上的客户端和服务器之间进行通信的接口(运行在 Windows95 或更高版本上)[4]。

本文提出的方法依赖于 ASP.NET 对 Word 程序的一系列操作，包括调用 Word 程序以及将 Word 文档另存为 XML 文件。首先，在 ASP.NET 中调用 Word 程序需要完成对 Word 对象库文件“MSWORD.OLB”的引用。由于安全原因，默认 ASP.NET 用户没有权限访问 Word.ApplicationClass()，因此需要在网站的 web.config 文件中添加代码 <identity impersonate="true"/>以启用模拟身份，这时所有网页将使用匿名 Internet 用户账户(IUSR_machinename)的权限执行任务。其次，通过对 DCOM 的设定来赋予该用户账户对 Word 程序的操作权限，主要包括在 Windows 的组件服务中对 Word 文档的 DCOM 接口进行配置，赋予该账户启动、激活和访问的权限，此时便可实现 ASP.NET 对 Word 程序的访问操作。

3　复杂格式数据传输的实现方法

本文所提方法的目的是实现 B/S 模式下用户在客户端进行复杂格式数据的提交及处理。由于 DSOFramer 控件能够嵌入 Word 应用程序，因此能够实现与 Word 程序相同的功能，故本文使用 DSOFramer 调用 Word 程序的方法来实现用户输入接口，同时使用户能够在熟悉的界面下进行操作。另外，该方法利用 Microsoft Office 对 XML 文档格式的支持，将 Word 文档另存为 XML 文档后转成文本数据，借助其标记的结构性，根据不同的标记取得 Word 文档中特定部分的内容，达到数据处理的目的，这也是实现此方法的关键。

3.1 实现方法的基本思想

用户在客户端使用 DSOFramer 控件调用 Word 程序对文档进行编辑，将文本信息以.doc 格式保存至服务器，由服务器将此 Word 文档另存为 XML 文件，通过对 XML 文件的结构分析提取与相应 Word 文档中的图文数据所对应的 XML 代码，将其以 XML 文本片段的方式保存到服务器，同时将该内容索引更新至 sql_server 数据库。用户进行操作时，通过对 XML 数据进行查找及合并等处理为用户提供查询等数据操作功能，最终生成报告反馈用户。实现过程如图 1 所示。

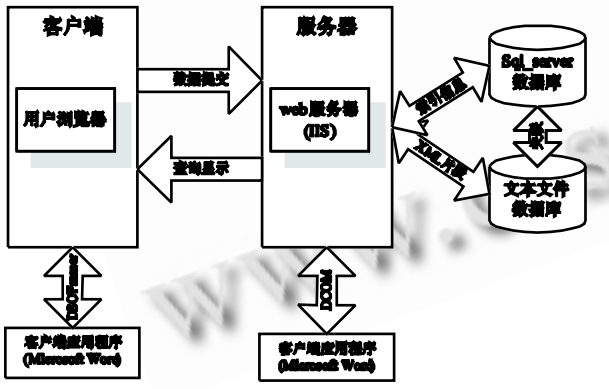

图 1　图文数据处理流程图

3.2 客户端文件上传的实现

此时该方法已通过 DSOFramer 控件使用户能够在客户端进行任意格式的输入，进而需要将用户输入的数据上传至服务器进一步处理。该操作使用 DSOFramer 自身的保存到服务器功能来实现，通过增加 HttpPost 上传接口，将用户提交的.doc 文件 Post 至一个动态页面(jsp,asp,php 等)，由动态页面负责解析数据[5]。本文所提出的方法将数据 Post 至 upload.aspx 页面，由该 ASP.NET 页面完成对数据的解析，并将此 Word 文档保存到服务器的特定目录下。

3.3 Word 数据转换为 XML 数据的方法

完成 Word 文档的保存后，使用程序中定义的 Word.Application 对象的 SaveAs()方法将此 Word 文档另存为 XML 文件，关键代码(VB 语言)如例程 2 所示：

例程 2　将 Word 数据另存为 XML

```
Dim wd As New Microsoft.Office.Interop.Word.Application
Dim owd As Microsoft.Office.Interop.Word.Document
owd = wd.Documents.Open(Server.MapPath("~/") &
"Word\temp.doc")
owd.SaveAs(Server.MapPath("~/") & "Word\temp.XML", 11)
```

(注：上述程序片断最后一行的“11”表示另存为的文件格式，这里即为 wdFormatXML 格式，同理如“8”表示 wdFormatHTML 格式，在客户端数据查询时会被使用。更多的另存为文件格式请参见 MSDN 的 Office Solutions Development。)

将 Word 文档另存为 XML 文档后，通过观察该文档的标记可以定位用户所输入数据的 XML 代码。使用 XML 对象定位所需的 XML 标记，提取相应代码保存到服务器上的文本文件中，并将相应的内容索引更新至 sql_server 数据库便于以后的数据查询。由于数据量较大，所以采用 txt 文件作为数据存储方式，并借助索引文件对数据库进行操作，而不是将 XML 代码直接保存到 sql_server 数据库中。

3.4 XML 数据的处理方法

由于 XML 格式具有良好的层次性，因而已保存的 XML 片段易于查询及合并，系统将用户输入的数据保存为 XML 文档后，就可以实现相应的数据操作，进一步组成新的 Word 文档。另外，通过程序中所定义的 Word 对象的 SaveAs()方法也可以将某部分特定内容另存为 HTML 文件返回给用户。由于索引信息存放在 sql_server 数据库中，通过对索引文件的操作能够轻松实现用户对于特定文档内容的查询，提高查询效率。

4　结论

本文提出的复杂格式数据传输的方法基于一个在

线文档生成系统的实现过程，用户要求采用分布式方法完成对复杂格式数据的传输，并对所提交的数据进行在线查询等操作，最终由系统生成 Word 格式或 HTML 格式的报告反馈用户。由于用户提交的数据量可能较大，并且格式复杂，因而提出了本文所述的解决方法。基于该方法实现的在线文档生成系统很好满足了用户需求且系统运行平稳。

该在线文档生成系统主要针对办公自动化及在网络环境下对复杂格式数据的提交和处理。本文提出的方法在系统实现过程中部署较为复杂，系统运行需要在客户端安装 Microsoft Word 程序，并注册 DSOFramer 控件来实现对 Word 程序的调用，即以降低客户端轻量级方面的要求为代价，来满足客户对复杂格式数据进行操作的需要。针对这一方法的不足之处，可以在该方法的实现思路上进行改进，借助在线文本编辑器 FCKeditor 直接进行简单格式图文信息的编辑、处理及传输，利用 XML 语言最终生成 Word 格式或 HTML 格式的报告。这种思路不仅可以达到客户端轻量级的目标，并且系统实现的过程更为简便，虽然无法很好地满足用户对复杂格式数据处理的需求，但仍具备在线文档生成系统的基本功能。基于这两种思路的在线文档生成系统均已实现。

参考文献

1 Visual C++ ActiveX Control for hosting Office documents in Visual Basic or HTML.2007 http://support.microsoft.com/default.aspx?scid=kb;en-us;311765#appliesto

2 W3C. Extensible markup language (XML)1.0[2004-2-4]. http://www.w3.org/TR/2004/REC-xml-20040204

3 杨新伦,唐培和,刘浩. ASP.NET 对 XML 文档的支持与处理方式.广西工学院学报, 2003,14(1):44－47.

4 崔应杰,张景,李军怀,孙东东,李朋.基于 XML 的 Web 系统.计算机工程, 2004,30(4):58－60.

5 张曜,张青. ASP.NET 函数实用手册.北京:冶金工业出版社, 2002,(12):88－91.