基于.NET 的移动话费详单页面的 XML 数据提取①

赵 纯，施一剑，张 昱，金心宇

(浙江大学 信息与电子工程学系，杭州 310027)

摘 要：介绍了一种在 Microsoft 的.NET 3.5 框架下，使用 ASP.NET、SgmlReader、LINQ 和 XML 等关键技术，对移动话费详单 HTML 页面进行自动数据提取的方案。该方案能实现对移动话费详单页面数据信息进行自动搜集的功能，有助于进一步完成对话费详单信息的统计、计算等处理工作。该方案具有简单、易行、高效的特点。

关键词：Web 数据提取；.NET 框架；XML；LINQ；话费详单

XML Data Extraction from Webpage of China Mobile Phone Bills Based on .NET

ZHAO Chun, SHI Yi-Jian, ZHANG Yu, JIN Xin-Yu

(Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: In this paper, a solution of automatic data extraction from HTML Webpages of China Mobile phone bills with the key technology such as ASP.NET, SgmlReader, LINQ, XML under the Microsoft .NET 3.5 framework is given. The solution provides the function of automatic data collecting from China Mobile phone bills, which contributes to a further process of the phone bills information. The solution is simple, feasible and effective.

Keywords: Web data extraction; .NET framework; XML; LINQ; phone bills

1 引言

中国移动为用户提供了多种多样的话费查询方式，如短信查询、营业厅查询、10086 客服电话查询以及互联网上在线查询。其中互联网上在线查询是一种非常方便的自助服务式的查询方法，也是用户进行话费详单查询的一种最常用的方法。这样查询得到的话费详单信息是以 HTML 页面方式呈现、以 HTML 标记语言描述的，虽然这种结构能够方便地呈现出所有话费详单信息，但是用户却难以用编程的方法对其中的数据信息进行自动搜集，因此不便于用户对于话费详单信息进行统计、计算等进一步的开发处理，从而难以准确、全面地了解话费详单的内容。

本文通过一个应用实例，介绍了一种在 Microsoft 的.NET 3.5 框架下，基于 Microsoft Visual Studio .NET 2008 开发平台，使用 ASP.NET、C#语言编程、LINQ(Language INtegrated Query)、XML（Extensible Markup Language）和 Microsoft SQL Server 等技术，对移动话费详单 HTML 页面进行自动数据提取的方案及其实现，该方案很好地解决了话费详单 HTML 页面数据难以编程提取并加以利用的问题。

2 在线HTML页面数据提取的方案

Web 网页一般使用 HTML 标记语言。HTML 语言比较擅长网页的布局和外观设置，但缺乏对网页内的数据信息的有机按序表达能力，而且 HTML 语言的语法要求也很不严谨，这使得想要通过编程方式直接从 HTML 页面中提取数据变得非常困难。随着 XML 技术的发展，W3C(World Wide Web Consortium)推出了 XHTML 标准。XHTML 是 XML 的一种应用，它既具有 HTML 语言擅长格式排版方面的特点，又使得设计人员能用 XML 编程方法来方便地处理其中的数据。本文介绍的在线 Web 数据提取方法就是先将现有的 Web 页通过转换工具 SgmlReader 从 HTML 格式转换成 XHTML 格式，再通过 XML 编程方法提取出其中

① 基金项目:浙江大学校级大学生科研训练计划(SRTP)第 12 期(4085)
收稿时间:2010-07-19;收到修改稿时间:2010-08-30

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的数据导入 SQL Server 数据库中以备进一步加以利用。本文中使用的 SgmlReader 来源于 Steve Bjorg 负责的基于 Microsoft .NET 框架的 SgmlReader 项目，2010 年 2 月 19 日发布的版本为 1.8.6[1]。

在数据提取方面，转换以后的 XHTML 文档比原来的 HTML 文档要容易许多。转换以后的 XHTML 文档中的内容很多都是网页格式和排版方面的信息，而需要提取的有用数据通常只集中在文档中的一小部分，一般都是包含在<table>等表格标记节点内的，这样就有利于将注意力集中在有用数据上。数据提取的关键是确定有用数据在文档中的位置，这可以借助于 XMPSpy[2]等专业软件，通过查看转换以后的 XHTML 文件，寻找有用数据所在位置以及包含这些数据的<table>标记节点的特点来实现[3]。

根据数据所在位置的特点，结合 XML 中的 XPath 和 XSLT 技术就能从 XHTML 文件中提取出所需数据并将其转换为 XML 文件。XPath 是 W3C 制定的对 XML 文件进行数据查询的规范，它是一种路径语言，通过 XPath 表达式说明到达所需查询节点的路径。XSLT 是 W3C 制定的转换 XML 文件的样式表规范，XSLT 文件中包含一组模板，XSLT 处理器根据该文件中的 XPath 表达式在对应的 XML 文档中寻找满足一定条件的节点与 XSLT 文件中的模板进行匹配，从而只取出所需的满足条件的节点。使用这两种技术，就能从 XHTML 文件中只提取出所需数据，而忽略掉所有其它无用信息。

这种方案的具体实现步骤如下：

(1) 确定 HTML 文档数据源并使用 SgmlReader 将其转换为 XHTML 文件；

(2) 确定所需有用数据在 XHTML 文件中的位置并设计提取数据使用的 XSLT 文件；

(3) 使用 XSLT 转换技术提取数据并将其保存为 XML 文件；

(4) 将转换得到的 XML 文件中的数据导入 SQL Server 数据库中；

(5) 利用 LINQ 技术对 SQL Server 数据库中数据进行处理分析。

其流程如图 1 所示。这种方案对 HTML 文档数据源有一定的要求，一般需要 HTML 格式的 Web 页面的结构相对稳定。

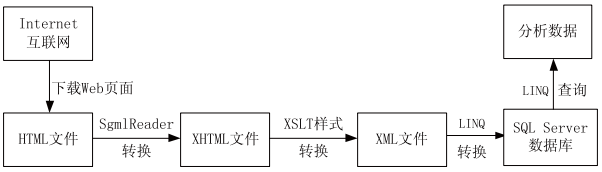

图 1　在线 HTML 页面数据提取方案流程图

3　移动话费详单页面数据提取方案及其实现

下面将依据上述在线 HTML 页面数据提取方案，设计程序来提取浙江移动公司网站（http://www.zj.chinamobile.com）上的用户通过在线查询自助服务而得到的话费详单 HTML 页面中的数据信息，并导入 SQL Server 数据库中以供进一步开发应用。这里以用户查询所得语音话费详单页面的数据提取为例，其页面示例如图 2 所示（用户已经将其保存到本地）。本文将提取出客户姓名及每条详单记录等信息，导入 SQL Server 数据库中，以便进一步进行一些数据统计和计算工作。


中国移动通信客户详单

图 2　浙江移动语音话费详单 HTML 页面

3.1 用 SgmlReader 将 HTML 转换为 XHTML

SgmlReader 是在 Microsoft .NET 框架下开发的用于解析 SGML 文件的工具，它也能解析 HTML 文件并将其转换为 XHTML 文件。SgmlReader 是从 Microsoft 的 XMLReader 类派生出来的，使用它解析 HTML 文件的用法与使用 XMLReader 解析 XML 文件是一样的[4,5]。在使用 SgmlReader 之前，必须将其以 DLL 格式封装的文件（SgmlReaderDll.dll）放入在 Visual Studio .NET 2008 平台下开发的 Web 应用程序解决方案的 bin 目录中，并在编程时导入其命名空间 Sgml。本文中采用 C#语言和 ASP.NET 开发，故需在

© 中国科学院软件研究所 http://www.c-s-a.org.cn

程序开头中加入语句“using Sgml;”。SgmlReader 输入的文件格式可以是 HTML 文件或其他 SGML 文件，文件可以位于本地硬盘上或远程服务器上，其输出为 XHTML 文件（XML 文件），默认编码格式为 UTF-8。

在本文中使用 SgmlReader 将浙江移动网站上用户通过自助服务查询得到的语音话费详单 HTML 页面文件转换为 XHTML 文件。首先将用户的语音详单 HTML 页面文件作为 SgmlReader 的输入项，然后通过使用.NET 中的 XMLTextWriter 类将 HTML 文件转换后写入到一个 XHTML 文件（XML 文件）中。部分 C#程序代码如下所示。

```
SgmlReader reader = new SgmlReader();     //定义 SgmlReader 类的一个实例
reader.DocType = "HTML";                  //指定 SgmlReader 类解析的文件类型为 HTML
reader.Href = filePath;                   //指定 SgmlReader 类解析的输入 HTML 文件
//定义 XMLTextWriter 类的一个实例，并将转换结果写入到输出 XHTML 文件（XML 文件）中
XmlTextWriter writer = new XmlTextWriter(filePath + ".xml", System.Text.Encoding.UTF8);
writer.Formatting = Formatting.Indented;
reader.WhitespaceHandling = WhitespaceHandling.None;
writer.WriteStartDocument();
while (!reader.EOF)
    if (reader.NodeType != XmlNodeType.Whitespace)
        writer.WriteNode(reader, true);
writer.WriteEndDocument();
writer.Flush();
writer.Close();
```

3.2 确定所需数据的位置，并设计提取数据使用的 XSLT 样式表文件

使用 SgmlReader 转换得到 XHTML 文件后，接下来就需要分析所需数据在 XHTML 文件中的位置。借助于可进行专业级 XML 应用开发的 Altova XMLSpy 2010[2]软件加以分析，可以发现转换得到的“语音详单.xml”文件的根元素是<html>，需要提取的客户姓名信息位于//html/body/form[1]/span[1]/table/tr[1]/td/table/tr[1]/td[1]节点内（其中的“[1]”表示第 1 个具有该名称的节点，如这里的“form[1]”表示父节点 body 下面的第 1 个 form 子节点），而需要提取的所有语音详单记录信息都位于//html/body/form[1]/span[1]/table/tr[2]/td/table/tr[2]/td/table/tr[3]/td/table/tr[@class!='listtitle']节点内（其中的“tr[@class!='listtitle']”表示其 class 属性不等于'listtitle'的节点 tr）。

在明确所需数据在 XHTML 文件中的位置后，可使用 XPath 和 XSLT 技术来设计提取所需数据的 XSLT 样式表文件。实际上，上述的节点位置表达式就是 XPath 表达式，利用这样的 XPath 表达式即可准确定位所需数据的位置。由于客户姓名数据在文件中是唯一的，因此可以直接利用上述 XPath 表达式来定位；而语音详单记录有很多条，需要借助 XSLT 提供的循环元素 for-each 去遍历每一条记录所在的节点加以定位。设计好的“语音详单.xsl”文件中的部分 XSLT 代码如下所示。

```
<xsl:template match="/">
<语音详单>
    <xsl:attribute name="客户姓名">
        <xsl:value-of select="normalize-space(substring(//html/body/form[1]/span[1]/table/tr[1]/td/table/tr[1]/td[1],6))"/>
    </xsl:attribute>
    <xsl:for-each select="//html/body/form[1]/span[1]/table/tr[2]/td/table/tr[2]/td/table/tr[3]/td/table/tr[@class!='listtitle']">
        <语音记录>
            <xsl:attribute name="序号">
                <xsl:value-of select="normalize-space(td[1])"/>
            </xsl:attribute>
            <通话起始时间><xsl:value-of select="normalize-space(td[2])"/></通话起始时间>
            <!-- 省略 -->
            <小计><xsl:value-of select="normalize-space(td[13])"/></小计>
        </语音记录>
    </xsl:for-each>
</语音详单>
</xsl:template>
```

3.3 使用 XSLT 样式表技术进行转换并生成 XML 文件

设计完成 XSLT 文件后，就可以进行实际的数据提取操作了。Microsoft 的.NET 3.5 框架提供了

© 中国科学院软件研究所 http://www.c-s-a.org.cn

XslCompiledTransform 类，可用来执行 XSLT 转换操作。XslCompiledTransform 类是.NET Framework 中支持 XSLT 1.0 语法的 XSLT 处理器。与过时的 XslTransform 类相比，该类是一个新的实现并且包括了性能的提升。其命名空间为 System.Xml.Xsl，程序集为 System.Xml（在 system.xml.dll 中）。XslCompiled Transform 类首先用 Load 方法载入 XSLT 文件，然后调用 Transform 方法实施转换。本文中采用的 Transform 方法的调用格式为：XslCompiled Transform.Transform (String 源 XML URI, String 目标 XML URI)，使用 URI 指定的输入文档执行转换，然后将结果输出到文件[6]。部分 C#程序代码如下所示。

```
string xslPath = Server.MapPath(@"Xsl\" + "语音详单.xslt");//指定转换所需的 XSLT 样式表文件
XslCompiledTransform xsldoc = new XslCompiledTransform();
xsldoc.Load(xslPath);    //载入并编译 xsl 样式表
/转换前的文件为：语音详单.xml，转换后的 XML 文件为：语音详单_data.xml
xsldoc.Transform(filePath + ".xml", filePath + "_data.xml");
```

3.4 将 XML 文件导入 SQL Server 数据库

完成 XSLT 转换后，语音详单 HTML 页面文件中的客户姓名及每一条详单记录等数据信息就都已经被提取出来并形成了一个单独的 XML 文件，最后再将这些提取出来的数据导入 SQL Server 数据库中以备进行后续的统计和计算等开发处理工作。这里采用 LINQ to XML 技术，该技术把 XML 文档放于内存中，使用 LINQ 对 XML 文档方便地进行查询 XML 节点、属性和值，转换 XML 树等[7]。LINQ 是一组技术的名称，这些技术建立在将查询功能直接集成到 C#语言的基础上。借助于 LINQ 技术，可以使用一种类似 SQL 的语法来查询任何形式的数据，并且代码少，效率高。目前为止 LINQ 所支持的数据源有 SQL Server、XML 以及内存中的数据集合等。

部分 C#程序代码如下所示。

```
MobileSystemDataContext database=new MobileSystemDataContext();//LINQ to Sql 数据库上下文
var phoneCallDetails = new List<PhoneCallDetail>();//保存所有详单信息的 List
XDocument xdoc=XDocument.Load(filePath + "_data.xml");//加载 xml 文件
foreach (XElement elem in xdoc.Descendants("通话详细记录"))//对于每一条通话详细记录
{
  var phoneCallDetail = new PhoneCallDetail()
  {
    序号=int.Parse(elem.Attribute("序号").Value.Trim()),//获取属性值
    通话日期起始时间 = DateTime.Parse(elem.Element("通话日期起始时间").Value),//获取子元素值
    通话时长 = int.Parse(elem.Element("通话时长").Value.Replace('秒', ' ').Trim()),
    //省略
    免费项 = elem.Element("免费项").Value.Trim(),
    网络类型 = elem.Element("网络类型").Value.Trim(),
  };
  phoneCallDetails.Add(phoneCallDetail);//加入 List 中
}
database.PhoneCallDetail.InsertAllOnSubmit(phoneCallDetails);
database.SubmitChanges();//提交更改，此时 SQL Server 才执行插入操作
```

这样，在 SQL Server 数据库的“语音详单”数据表中，就导入了最初的用户查询语音详单 HTML 页面文件中的客户姓名及每一条语音详单记录数据了。该数据表的部分内容如图 3 所示。

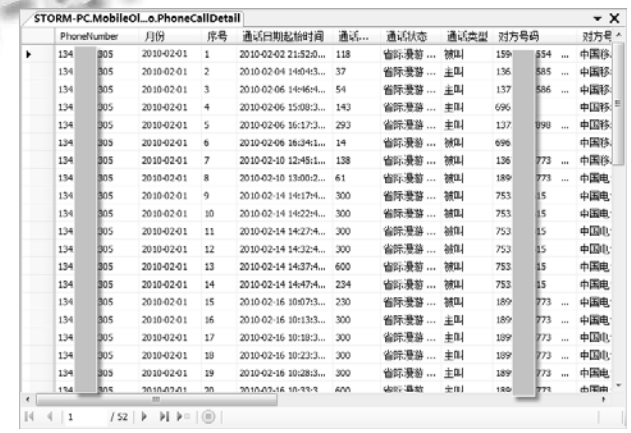

图 3　最后导入的 SQL Server 数据表

之后就可以进一步利用 C#语言和 LINQ to SQL 技

© 中国科学院软件研究所 http://www.c-s-a.org.cn

术，处理 SQL Server 数据表中的数据。LINQ to SQL 是.NET Framework 3.5 版的一个组件，提供了用于将关系数据作为对象管理的基础结构。本文中运用 LINQ to SQL 技术的部分 C#程序代码如下所示：

```
using System.Linq;
//…代码省略
int monthcount = monthList.Count();
for (var i = 0; i < monthcount; i++)
{//查询用户几个月的通话时间
var CallTime = from n in db.PhoneCallSum
where n.PhoneNumber == phoneNumberDropDown
List.SelectedValue && n.查询周期==monthList[i]  select n;
//…代码省略
}
```

4 结束语

本文针对中国移动网上话费详单页面中的数据难以进行提取处理的问题，介绍了一种在 Microsoft 的.NET 3.5 框架下，使用 ASP.NET、SgmlReader、LINQ 和 XML 等关键技术，对移动话费详单 HTML 页面进行自动数据提取的方案。该方案具有简单、易行、高效的特点。通过这一方案能够实现对移动话费详单页面数据信息进行自动搜集的功能，并使用户能进一步完成对话费详单信息进行统计、计算等开发处理工作，从而可使用户准确、全面地了解话费详单情况。该方案已经成功应用于浙江移动即将推出的“移动套餐专家”系统中。

参考文献

1 Bjorg S, Touch M. Browse MindTouch (frmly deki wiki) Files on SourceForge.net (SgmlReader Download Page) [2010-03-01]. http://sourceforge.net/projects/dekiwiki/files/SgmlReader, 2010

2 Altova Company. XML Editor, Data Management, UML, and Web Services Tools from Altova [2010-04-01]. http://www.altova.com/

3 Jackson J, Myllymaki J. Web-Based Data Mining (from IBM developerWorks)[2010-03-15]. http://www.myllymaki.org/jussi/webdata/dworks2001.pdf

4 Wahlin D. Parse HTML Pages to Extract Data [2010-02-15].http://www.ftponline.com/xmlmag/2002_12/online/xml_wahlin_12_18_02/default_pf.aspx

5 耿建勇,鲁士文.微软.NET 框架下提取在线 Web 数据的方法.计算机系统应用,2004,13(4):53–56.

6 Wahlin D.王宝良译.基于 XML 的 ASP.NET 开发.北京:清华大学出版社,2002.

7 龚赤兵.Visual Studio 2008 中的 LINQ 开发技术.北京:机械工业出版社,2008.

（上接第 222 页）

用，相关研究还在不断深入，这些都为构建高质量大规模的应用系统提供了强有力的支持。

参考文献

1 张驰.基于接口匹配的构件组装.计算机应用,2007,27(6):1420–1422.

2 张世琨,张文娟,常欣,王立福,杨芙清.基于软件体系结构的可复用构件制作和组装.软件学报,2001,12(9):1351–1359.

3 孙莹,陈松乔.接口连接式构件组装的一种形式化方法.计算机科学,2006,33(7):253–256.

4 任洪敏,钱乐秋.构件组装及形式化推导研究.软件学报,2003,14(6):1066–1074.

5 毛莺池,梁奕,王志坚.异构软件构件组装模型设计与实现.计算机工程,2005,31(7):56–58.

6 陈建勋,张子鹤,马于涛,夏学知.基于软件体系结构的组件集成框架.计算机工程,2004,30(22):68–71.

7 张景山,廖华明,侯紫峰,徐志伟.普及计算中基于接口语义描述的动态服务组合方法.计算机研究与发展,2004,41(7):1124–1134.

8 孟凡超,初佃辉,战德臣.基于行为的构件组合及存在性检查.计算机工程,2009,35(15):41–44.

9 黄万艮,陈松乔.基于消息的构件组合运算与构件演化.小型微型计算机系统,2007,28(7):1216–1220.

10 黄靖,卢炎生,徐丽萍.适时构件合成的语义研究.计算机科学,2007,34(2):10–16.

© 中国科学院软件研究所 http://www.c-s-a.org.cn