

OpenOffice.org 文档结构的研究与分析

李映, 朱斐

(苏州大学 计算机学院, 江苏 苏州 215006)

摘要: 阐述 OpenOffice.org 文档的压缩存储方式以及这种方式和其他的多种存储方式相比的优势; 针对 OpenOffice.org 的各种文档描述了它们的结构, 并且讨论了从 OpenOffice.org 的 XML 文档中读取元数据方法; 最后进一步讨论了 OpenOffice.org 的 XML 元数据读取的意义以及它和文档结构化其他研究的关系。

关键词: 压缩存储; XML; 格式化文档; 文档结构; OpenOffice.org

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2006)10-0058-04

Research and Analysis of OpenOffice.org Document Structure

LI Ying, ZHU Fei

(Computer School of Soochow University, Suzhou 215006, China)

Abstract: Explains the compression storage of the OpenOffice.org document and advantage of this means comparing to other ones. Describes the document structure in the OpenOffice.org and discusses the methods of reading metadata from OpenOffice.org XML document. Finally more discussion about its role to other research is presented.

Key words: compression storage; XML; formatted document; document structure; OpenOffice.org

0 引言

在信息时代, 文档仍是信息交换的主要媒介。随着人们工作规模和复杂程度的提高, 各种文档的之间的兼容性已经成为迫切需要解决的问题。不同的应用软件对文档都按照各自定义适用的格式存储。为了避免这种混乱状态, 世界组织都提出了文档结构化。所谓文档结构化就是用一种明确的、无二义性的方式组织文档的各个单独部分。OpenOffice.org 就是文档结构化领域的一个典例应用。

1 文档的存储格式

为了解决文档格式不兼容和相互封闭性的问题, OASIS 定义了开放办公软件文档格式规范——OpenDocument^[1]。OpenDocument 已通过 OASIS 标准投票, 成为欧盟政府推荐规范, 美国马萨诸塞等州政府选定 OpenDocument 为标准文档格式, IBM 和 Sun 联手支持 OpenDocument 开放规范, KDE 环境下的 Koffice 及 Corel WordPerfect 也都将对其提供支持。OASIS 组织将 OpenDocument 格式提交给 ISO(国际标准化组织)进行标准认证, 在通过认证之后, OpenDocument 将正式成为开放性的公有文档格式标准。OpenDocument 得到了广泛和有力的支持, 同时 OpenDocument 的开放性能很好地满足中国软件自主

产权的计划, 也使得文档的共享更为容易。

OpenDocument 是新一代的办公文档格式, 提供了包括字处理、电子表格、演示软件等桌面软件用于存储文件的格式。OpenDocument 是一个开放的标准, 它的发展将打破 MS Office 的封闭办公文档格式的约束, 有利于文档的交流。

目前, 国内广大的办公软件市场都将被国外知识产权的软件占领, 不仅是产业发展受到影响, 而且在资料的数字化存储、管理和安全控制上无法得到保证。在这种形势下, 国家推动了从操作系统到应用软件的整体规划。在操作系统层次上, 推动基于 Linux 核心的、具有独立知识产权的操作系统; 在应用软件层次上, 提出了支持包括办公软件、浏览器、个人信息助理等软件产品。OpenDocument 格式以其统一性、开放性、兼容性和规范性受到了国外 Sun, IBM 等公司, 国内金山、永中、中文 2000、京华网络等公司的支持。在 2002 年的政府采购中, 国产软件包括金山、中文 2000 等公司拿到了超过一半份额的订单, 而且随着永中 Office 的逐渐成熟, 电子政务中使用的国产中文办公软件将会进一步增加比重, OpenDocument 格式的应用领域也将越来越广。

OpenOffice.org 采用 OpenDocument 规范作为标准文档格式, 是全球三大开源项目之一。Sun 公司的 StarOffice 8 软件、Google 和 Sun 公司计划中的 WEB Office 产品、金山、永中、中文 2000、京华网络等其他国内公司都提供了对 OpenOffice.org 文档的支持。表 1 列出了 OpenOffice.

收稿日期: 2006-01-02

作者简介: 李映(1976-), 女, 江苏苏州人, 硕士研究生, 工程师, 研究方向 MIS、网络与数据库。

org 默认文档类型^[2]。

表 1 OpenOffice.org 的常见文档类型

应用程序名称	文件扩展名
OpenOffice.org Writer	*.sxw
OpenOffice.org Writer 模板	*.stw
OpenOffice.org Calc	*.sxc
OpenOffice.org Calc 模板	*.stc
OpenOffice.org Impress	*.sxi
OpenOffice.org Impress 模板	*.sti
OpenOffice.org Draw	*.sxd
OpenOffice.org Draw 模板	*.std
OpenOffice.org Math	*.sxm
主控文件	*.sxs

过去,系统没有提供足够的内、磁盘空间和 CPU 性能,所有存值的不得以复杂的文档结构存储,即以 ASCII 码存储文档内容,而文本本身的布局设计和风格等信息则以二进制代码记录下来。这样各自的应用程序有相应的文档文件。文件的格式与它们各自的应用密切相关,彼此互不兼容,而且随着应用程序的版本更新,新的要求需要新的数据格式和新的格式。这就导致同一应用程序不同版本之间的不兼容,使得人们不得不用同一程序同一版本的软件,给人们之间的文档交换造成了困难。

另一种是将样式与内容分离开来表示,通过某种映射联系起来。OpenOffice.org 支持的 XML 文件格式采用这种实现方式,XML^[3]存储的优点主要有:公开的、统一的、多用途的文档格式;文件小;内容数据和格式分离;XML 被广泛地支持;XSLT 的 XML 工具可以对 XML 文档进行协同;第三方的支持,大大地便于人们分离内容和描述信息。同时,将文档存储为 XML 文档格式,使得 OpenOffice.org 做到了真正完全的公开,除了开放源代码,还可以不通过 OpenOffice.org,而通过程序就可以获得 OpenOffice.org 格式化的文档。这是 MS-Document 做不到的。更为重要的是 XML 允许用简单、免费的小型工具就可以对 OpenOffice.org 文档进行各种操作。

2 文档结构和内容

OpenOffice.org 的上述文件其实就是 ZIP 格式的压缩文件。正是因为这个原因使得打开 OpenOffice.org 文档比较慢,可以看到打开的进度条展开的过程。这主要是在解析 XML 的压缩文件。另一方面,当把 OpenOffice.org 文档的扩展名强制更改为 *.zip 内容格式的文件时,可以得到如下的文件和内容^[1]。

(1)meta.xml:文件的属性,包括一系列具有文档元数据的元素(如作者创建和最后编辑日期、已经花费在编辑该文档上的总时间、字数、页数、表数和图数等元素)。如果以密码存储文件,则只有 meta.xml 不会被加密。

(2)styles.xml:文件中使用的 styles 可以将 styles.xml

看作 XML 格式中的级联样式表(CSS)和 XSL 格式化对象(XSL-Formatting Objects, XSL-FO)之间的交叉点。它定义了各种样式,这些样式可用于文档的字体、间距、修饰、间隔、制表符停止位等方面的编辑会话。它命名了所有样式,因此可以在其它文件中引用它们。

(3)content.xml:文件本文内容(文字、表格、图形元素),这是文档内容的核心部分。

(4)settings.xml:文件和浏览相关参数。如放大比例及设定的印表机选项、记录的变更、相关资料等。这些参数通常是由应用程序预先设定,记录用户对 OpenOffice.org 用户界面的用户首选项。这些涉及到用来编辑文档的应用程序的细节,而不是文档本身的任何细节。这个领域还需要完成一些工作以确保互操作性。

(5)Mimetype:记录 XML 文档对应的解析程序。如 application/vnd.sun.xml.writer; application/vnd.sun.xml.calc。

(6)Layout-cache:文件的子目录。

(7)META-INF/manifest.xml:整个文档的 XML 文件结构信息,指出每个文件的存储类型以及提供了当文档加密时,解密的方法。

(8)Pictures/:存放图形的目录(以原始二进制格式储存如 png 图像或者 JPEG 图像)

(9)Dialogs/:存放宏定义时,所使用的对话框。

(10)Basic/:存放以 StarBasic 撰写的相关脚本信息。在它的子目录 Standard/StarBasic 语言编写的可执行的脚本。子目录 Script.xml 给出库文件的名称。

(11)Object/:用来存放嵌入式对象(如图表);每一个目录都含有一个对象,以独自的原始格式储存。对于 OpenOffice.org 的物件,通常以 XML 的表示方法储存,而对于其它对象则通常会以二进制格式来储存。

3 对 OpenOffice.org 文档的提取

除了 OpenOffice.org 对 OpenOffice.org 文档格式文档的内容读取之外,也可以通过编写程序的方法对 OpenOffice.org 文档内容进行提取,主要的方法有如下几种。

3.1 Java 程序对 OpenOffice.org 文档内容的提取

利用 XML 的分析器(XML Parser)对 OpenOffice.org 文档进行解析,然后就可以提取其中相关的内容。首先,要获得 SAX(Simple API for XML)的相关包或者 DOM(Document Object Model)^[4]。以下是给出的部分源程序:

```
.....
import org.xml.sax.Attributes;
import org.xml.sax.ContentHandler;
import org.xml.sax.ErrorHandler;
import org.xml.sax.Locator;
import org.xml.sax.SAXException;
import org.xml.sax.SAXParseException;
import org.xml.sax.XMLReader;
import org.xml.sax.helpers.XMLReaderFactory;
```

```

.....
XMLReader parser = XMLReaderFactory. createXMLReader
("org.apache.xerces.parsers.SAXParser");
//注册自己设计的内容处理器
parser.setContentHandler(new MyContentHandler());
//注册我们的错误处理器
parser.setErrorHandler(new MyErrorHandler());
//分析文件
parser.parse(uri);
.....//自己设计的内容处理器,重载 SAX 中的 ContentHandler 类方法;
public void startDocument ()
public void endDocument()
public void startPrefixMapping (String prefix, String uri)
public void endPrefixMapping (String prefix)
public void startElement (String namespaceURI, String local-
Name,String qName, Attributes atts)
public void endElement (String namespaceURI, String local-
Name,String qName)
public void characters (char ch[], int start, int length)
public void ignorableWhitespace (char ch[], int start, int
length)
public void skippedEntity (String name)

```

在对 OpenOffice.org 文档进行解析时应当给出 OpenOffice.org 的 DTD, 因为 OpenOffice.org 对文档解析要按照 DTD^[3]的格式进行分析。

3.2 Perl 程序对 OpenOffice.org 文档内容的提取

Perl^[5]对 OpenOffice.org 文档的解析需要库 XML::XPath^[3]的支持。以下给出部分源程序:

```

use XML::XPath; //引入 XML 的分析库文件
while (my $file = shift @ARGV) { //逐个得到参数中的文件名
    next unless -f $file; //无效文件需剔除
    eval {
        my $xp = XML::XPath->new(filename =>
$ file); //生成一个 XML 对象
        print $S_ , " : ", $xp->findvalue("//dc:cre-
ator"), "\n"; //对 XML 对象提取相关元素
    };
}

```

3.3 XSLT 对 OpenOffice.org 文档内容的重构

XSLT 用于描述从 XML 文档到任意基于文本的格式的转换^[6]。它通过一系列的 XPath, 在转换过程中生成新的 XML 文档。它允许动态生成的元素和属性的自由的加入, 从一个 DTD 到另一个 DTD 的转化, 以及不再需要的信息的删除。常用的 XSLT 元素有:

- * <xsl: element>, 用于产生新的 XML 元素。
- * <xsl: attribute>, 用于产生新的 XML 属性。
- * <xsl: attribute-set>, 用于产生一组新的 XML 元素。
- * <xsl: text>, 用于产生文本。

但是 XSLT 并不是一种编程语言。在多数的团体中,

大量重要的数据被存储在关系型数据库中^[6]。XSLT 并未提供通过样式表直接访问这些数据的方法; 样式表开发人员必须使用数据提取工具或者自编程序来获得需要处理的数据。

3.4 ASP, JSP 等程序对 OpenOffice.org 文档内容的提取

这是一种广为使用的方法, 这里只作简要的说明。以下是 ASP 对 content.xml 文档中文本内容提取的片断。

```

<%@ LANGUAGE="VBScript" %> //标签标识的语言
为 VBScript

```

```

.....
Set parse = server.CreateObject("MSXML2.DOMDocument")
//创建一分析器
Parser.async = false
Parser.load(Server.MapPath("content.xml")) //装载一个
XML 文档

```

```

.....
Set n1 = parser.selectNodes("//text:p") //获得 XML 文档中
含有<text:p>元数据的结点列表
For i = 0 to n1.length - 1
    Response.Write "<tr><td>" //将文本<tr><td>发送
到客户端

```

```

Response.Writer n1.item(i).text //输出文本内容

```

```

Response.Write "<tr><td>"

```

```

next

```

以上只是讨论了 4 种方法, 其它的语言, 如 Visual C++ , Visual Basic 等也提供了对 OpenOffice.org 文档的解析方法, 限于文章篇幅, 在此不一一讨论了。

4 OpenOffice.org 文档 XML 元数据

OpenOffice.org 的文档是多个 XML 文件经过 ZIP 压缩方法压缩后的一个文件包。可在 {installpath} \ share \ dtd 中找到 DTD (Document Type Description) 文件。以下是对 OpenOffice.org 文档中的 XML 元数据的分析。

4.1 meta.xml 中的元数据含义

它的标签定义隶属于命名空间: http://openoffice.org/2000/office 下的 office: meta.^[1]。在 OpenOffice.org 中常用的标签有:

Generator: 文档的生成编辑器的名称

Title: 作品的名称

Description: 作品的摘要或者作品的内容叙述

Subject: 作品的主题

Keywords: 作品的关键字

Initial Creator: 作品的前期的筹备人

Creator: 作品的创作者或者组织

Printed By: 文档的打印设备的名称

Creation Date and Time: 作品的创作日期和时间

Modification Date and Time: 作品修改的日期和时间

Print Date and Time: 文档打印的日期和时间

Document Template: 作品使用的文档类型

Automatic Reload;

Hyperlink Behavior;

Language:作品本身使用的语言

Editing Cycles:作品的修改周期

Editing Duration:作品所涵盖的时间

Document Statistics:文档的统计数据,包括字数统计、单词数统计、段数统计、页数统计、内嵌对象数统计、图片数统计、表格数统计。

4.2 Content.xml 中的元数据含义

Text 中的元数据含义^[1]:

①p:文档中的一个段落。它是 text 中最为重要的标签。这里简要说明它的基本属性:style-name:给出这个段落的字体显示的风格;cond-style-name:给出在条件满足时,段落的字体显示的风格。

②h:指出 OpenOffice.org XML 文档的篇章结构。它的基本属性:style-name:给出这个段落的字体显示的风格;cond-style-name:给出在条件满足时,段落的字体显示的风格;level:层次。

③span:代表段落中部分文档所使用的某一风格的范围。它的基本属性:style-names。

④s:代表空格(space)的数目。

⑤a:代表文档中含有超链接数据元素:它的基本属性有:type:超链接的类型,如 simple。

⑥href:超链接的 URL 地址;target-frame-name:超链接文档的显示框体;-self:用超链接文档取代当前文

档;-blank:在一个新的框体中显示超链接文档发,parent:在当前文档的父类框体中显示超链接文档;-top:在当前文档的最高层父类框体中显示超链接文档。

5 小 结

文中比较详细地介绍了 OpenOffice.org 的文档的存储格式以及在文档解压后的文档结构。并且针对其的 XML 文档内容用多种方法进行内容的提取,详细讨论了文档结构的元数据,希望能对文档结构化、文档的协同粒度的降低提供帮助。

参考文献:

- [1] MICHAEL. Open Document Format for Office Application v1.0[EB/OL]. www.oasis-open.org. 2005-05-01.
- [2] Perry G. Sams Teach Yourself OpenOffice.org 2. Sams; Bk&CD-Rom edition[Z]. 2005.
- [3] MCLAUGHLIN B. JAVA 与 XML(第 2 版)[M]. 北京:中国电力出版社,2004.
- [4] HORSTMANN C S, CORNELL G. Core Java 2, Volume II - Advanced Features (7th Edition)[M]. 北京:机械工业出版社,2006.
- [5] BROWN M C. XML 与 PERL、PYTHON 和 PHP 编程指南[M]. 北京:电子工业出版社,2004.
- [6] GARDNER J R, RENDON Z L. XSLT 和 XPATH - XML 转换指南[M]. 北京:机械工业出版社,2002.

(上接第 57 页)

行更新处理,否则必须重新命名。首先读模型文件并检验其正确性,导入数据库,然后导入数据文件,检验正确性,在数据库中进行检验是否插入。导入程序流程如图 3 所示。

CIM 强调的就是互操作,在 DTS 数据库基础上,系统设计了基于 CIM 的自动导入\导出程序,成功地实现了 CIM\XML 文件的互导。为了检验第三方系统的信息交换,从 EMS 系统导出实时断面,转化为 CIM 格式文件,导入 DTS 系统,进行潮流计算,计算结果符合电力系统模型。

4 结束语

从长远来看,实现不同电力系统软件间的“即插即用”是必然的发展方向,深入理解并跟踪 CIM 标准的制定,对于中国电力行业的标准化有重大意义,更利于电力产品的国际合作。文中在这一基础上重点阐述了将 XML 技术应用于 CIM 的规则与设计流程,并结合实际开发中的 DTS 项目,实现了 DTS 数据库系统在 CIM\XML 格式下的导入导出,验证了与 EMS 系统进行数据交换的可行性,为下一步的工作打下了基础。

参考文献:

- [1] IEC61970. Energy Management System Application Program Interface (EMS-API) Part 301: Common Information Model (CIM) Base Draft Revision5[S]. 1999.
- [2] IEC61970. Energy Management System Application Program Interface (EMS-API) Part 501: CIM RDF Schema Draft Revision2[S]. 1999.
- [3] 辛耀中. 新世纪电网调度自动化技术发展趋势[J]. 电网技术, 2001, 25(12): 100-101.
- [4] 张慎明, 刘国定. IEC 61970 标准系列简介[J]. 电力系统自动化, 2002, 26(14): 1-6.
- [5] deVos A, Widergren S E, Zhu J. XML for CIM model exchange[A]. PICA 2001, IEEE Conference[C]. Sydney, Australia: [s. n.], 2001.
- [6] 刘崇茹, 孙宏斌, 张伯明, 等. 基于 CIM XML 电网模型的互操作研究[J]. 电力系统自动化, 2003, 27(14): 45-48.
- [7] 吴文传, 张伯明, 徐春晖. 调度自动化系统实时数据库模型的研究与实现[J]. 电网技术, 2001, 25(9): 28-32.
- [8] 鲁杰爽, 石东源. 基于 CORBA/XML 的电力企业应用集成[J]. 继电器, 2003, 31(12): 29-32.