ISSN 1009-3044
Computer Knowledge And Technology 电脑知识与技术
Vol.3,No.6,August 2008, pp.1338-1340

E-mail: kfyj@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

XML 技术在 NMR 系统软件参数处理中的应用

张太彪[1,2,3], 曾文华[2,3]

(1.厦门工商旅游学校 计算机教研室, 福建 厦门, 361000; 2. 厦门大学软件学院, 福建 厦门, 361005; 3. 智能信息技术福建省重点实验室, 福建 厦门, 361005)

摘要: 本文针对 NMR(核磁共振)系统软件的需求为其参数文件提供了一种 XML 设计方案, 通过对多种 XML 解析方案的实验和比较, 为 XML 参数文件提供了一种最优的解析和处理方案(VTD-XML), 并且论证了 VTD-XML 在大数据处理方面的优势。

关键词: VTD; LC; SAX; 文档对象模型; 核磁共振; 非提取式解析

中图分类号: TP391 **文献标识码**: A **文章编号**: 1009-3044(2008)24-1338-03

Application of XML in Parameters Processing of Software for NMR

ZHANG Tai-biao[1,2,3], ZENG Wen-hua[2,3]

(1.Computer Dept. of Xiamen Commercial School, Xiamen Fujian, 361000, China; 2. Software School of Xiamen University, Xiamen Fujian, 361005, China; 3. Key Laboratory for Intelligent Information Technology of Fujian Province, Xiamen University, Xiamen Fujian 361005, China)

Abstract: According to the commands of parameters processing in software for NMR, the authors provided an approach with XML for the parameters file, and brought forth a best solution (VTD-XML) for XML processing by comparing it with others techniques of XML parsing. Lastly they draws a conclusion that VTD-XML is more appropriate for large XML document than others similar techniques.

Key words: VTD; LC; SAX; DOM; NMR; non-extrative

1 XML 技术简介

近年来, XML(Extensible Markup Language, 可扩展标记语言)技术已经成为信息技术中最引人注目的技术之一。作为新一代标记语言, XML 凭借其众多优势, 已经逐步成为了数据表示的一个开放标准, 在许多领域得到了广泛的应用。

XML 语言的前身是 SGML 和 HTML, 它们是两种非常成功的标记语言, 但都在某些方面存在着技术缺陷[1]。与它们相比, XML 具有以下显著的特点[2]: (1)可扩展性, XML 保留了 SGML 的可扩展功能, 允许不同的专业开发与自己特点领域有关的标记语言; (2)自描述性, XML 文档通常会包含一个文档类型声明, 因此它能够被方便地交换和处理; (3)结构化和集成的数据, XML 能够表现许多复杂的数据关系, 并能使不同来源的结构化数据轻易地结合在一起; (4)数据和数据的显示方式实现分离, XML 的数据存储格式不受显示格式的制约, 它提供了一种结构化的数据表示, 使数据和显示分离。如今, 由于 XML 各种良好的特性, 越来越多的数据以 XML 格式保存和传输, 然后再根据不同的应用场合把 XML 转换成相应的数据格式, 很好地解决了异构数据之间的转换问题。

2 NMR 系统软件中参数文件的 XML 设计方案

2.1 NMR 系统软件开发背景及功能介绍

磁共振成像(MRI)技术作为一种能反映多维信息的无损伤诊断手段, 在医学病理和基础科学研究方面得到了广泛的应用。目前国内所生产的低场永磁 MRI 系统中, 绝大部分主要部件都依赖于进口, 尤其是谱仪部件[3]。磁共振厂商的竞争越来越体现为谱仪技术的竞争。为了推进谱仪的国产化, 国家投入了大量的资金用于数字化谱仪的研究(本项目并是在这样的背景下展开的), 研究的重点包括软硬件两方面的开发。MRI 谱仪是 MRI 系统硬件的核心部件, 主要应用于开展小型磁共振成像系统的应用和研究, 本文所指的 NMR 系统软件属于这一项目的软件分支。

NMR 系统软件是一套实现数字化谱仪控制的计算机软件系统。系统主要有四大功能模块: 仪器控制模块、实验设计模块、数据处理分析模块和系统管理模块。其中, 仪器控制模块完成数字化谱仪的硬件控制, 以及为其他模块提供与仪器通讯的编程接口; 实验设计模块为研究人员或脉冲序列开发人员提供一个功能强大、易于使用的脉冲系列编程环境; 数据处理分析模块负责对 1D 和 2D 数据的处理、显示和保存功能, 便于数据的分析和处理; 系统管理模块主要完成系统用户管理、机时管理及文件管理等有关功能。

2.2 参数文件的 XML 设计方案

参数作为贯彻整个 NMR 系统软件的主线, 在系统的各功能模块中起着非常重要的作用。换句话说, 参数设计方案的优劣在很大程度上成为系统性能好坏的一个主要的标准。

为克服现有软件的不足之处, 推进数字化谱仪的国产化, 研发出与目前世界水平相当的谱仪软件系统, 我们经过认真的考虑, 最后决定使用 XML 作为系统的参数设计方案。因为 XML 不仅具备了平台无关的特性, 还可以实现高度结构化且与显示无关的数

收稿日期: 2008-06-19
基金项目: 国家科技支撑计划课题(2006BAK03A22)
作者简介: 张太彪(1980-), 男, 福建三明人, 厦门工商旅游学校计算机教师, 硕士研究生, 主要研究方向为软件工程、仪器软件开发(taibiao@gmail.com); 曾文华(1964-), 男, 江苏兴化人, 教授, 博导, 博士, 主要研究方向为人工智能、软计算、网格计算、计算机体系结构、嵌入式系统、智能控制。

据描述。我们采用参数的 XML 分层描述，将实验中各 block 的公共参数放在顶层，而与各 block 相关的参数放入单独的一层中，与标记有关的参数也单独存放，以实现参数的分层管理。在实际使用中，至少会有一层，对于一个较复杂的实验可能需要用几千层来描述参数。

参数文件一般由三个部分组成，分别为公共参数部分(common)、层参数(layer)和标记参数(annotation)，其中的每一部分均有多个参数。通过它这种设计可以满足系统的相关需求，实现复杂实验的参数管理。

3 XML 文件的解析及其方法

XML 只是一种以纯文本对数据进行编码的格式，要想利用 XML 文件中所编码的数据，必须将其从 XML 纯文本中解析出来，因此，必须有一个能够识别 XML 文档信息的文本文件阅读器(即 XML 解析器)，用于解释 XML 文档并提取其中的内容[1]。目前 XML 的解析模式主要有两种：提取式解析(extractive parsing)模式和非提取式解析(no- extractive parsing)模式[2]。提取式 XML 解析在解析 XML 文档时，提取一部分原文件，一般是一个字符串，然后在内存中对其进行对象的构建。传统的 XML 解析均是采用这种方式，其主要的代表是 SAX(Simple API for XML)和 DOM(Document Object Model)。非提取式 XML 解析不同于前者，它在解析 XML 时，将文档作为整体一次性读入，然后以二进制数组的形式处理 XML 数据。这种解析模式的代表是 VTD- XML(Virtual Token Description for Extensible Markup Language)。下面具体讨论一下这几种 XML 解析技术。

3.1 文档对象模型 DOM

DOM 是由 W3C 制定的一种独立与语言和平台的标准，其实质是一组用于定义、创建和处理 XML 文档结构及其内容的 API，它提供了一个可以通用于各种程序语言、操作系统和应用程序的接口。

DOM 是一种基于树结构的 XML 解析技术，它将结构完整的 XML 文档定义为一棵树，开发人员只需利用树中的对象便可以轻松对 XML 文档进行读取、搜索、修改、添加和删除等操作。标准的 W3C 中 DOM 有这几种节点：Document、Root、Text、Element 和 Attribute。使用 DOM 对 XML 文件进行处理解析时，它将文档中的元素、属性、注释和处理指令等都看作节点(Node)，然后在内存中以节点树的形式创建 XML 的文件表示[4]。这样便可以通过节点树来访问 XML 文档中的内容，并根据需要来修改文档，以上过程便是 DOM 的基本工作原理。

DOM 解析技术的最大优点是简单易用且方便编程，因为当 XML 文档解析完毕后，整个文档并以一棵文档树的形式被保存在内存当中，操作非常方便，支持删除、修改和重新排列等众多功能。但是，基于 DOM 的解析程序运行效率不高，且消耗内存极大，不大适合大型 XML 文档的解析处理。

3.2 简单应用程序接口 SAX

与 DOM 不同的是，SAX 是一种基于事件的 XML 解析技术，它不是 W3C 提出的标准，但是功能很强大，因此在实际中应用相当广泛，几乎所有的解析器都会对它支持。

SAX 提供了一种对 XML 文档进行顺序访问的模式，这是一种快速读取 XML 数据的方式[5]。当使用 SAX 对 XML 文档进行解析时，解析器会从头到尾将 XML 文件顺序读入，读入过程中每遇到相应部分并会产生具体的事件，然后应用程序再根据事件的发生来调用相应的事件处理方法，完成每个事件的处理。例如，当 SAX 遇到某一 XML 元素的开始标记时，它会触发一个元素开始(StartElement)事件，并在事件中提供相应的信息，如：元素的名称、属性集等等。当然，前提是程序员必须在此过程中对自己感兴趣的事件进行相应的编码。SAX 类似于流媒体处理，分析能够立即开始，应用程序只是在读取数据时检查数据，因此不必将数据存储在内存中(这一点是 DOM 没办法做到的)。

3.3 XML 快速解析技术——VTD- XML

前面介绍的两种 XML 解析技术均属于提取式解析，解析过程中伴随着大量对象的创建，效率相对低下，对内存的消耗较大。为了克服提取式解析的缺陷，一种全新的 XML 解析模式—非提取式解析，意蕴而生。非提取式解析在解析 XML 文档时，将文档以二进制的形式原封不动地读入内存，以二进制数组保存文档信息，通过这一数组来获取、修改和更新 XML 文档中的数据，处理效率非常高，内存占用只比 XML 文件本身大不了多少。

非提取式解析模式的一个典型代表是 VTD- XML，它在 XML 数据加载时不做任何的解码工作，把 XML 的基本结构以二进制形式读入内存，并对二进制数组加以解析，获得元素的位置信息，将其存入位置缓冲器 LC(Location Cache)中，然后便能以数组的访问效率来获取数据，其访问的复杂度为 O(1)。

为了让大家进一步了解 VTD- XML，下面介绍两个重要概念：VTD 和 LC。需要注意的是 VTD 并非一个 API 规范，它只是关于如何编码令牌中各种参数的二进制格式说明。当前 VTD- XML 模型中采用的 VTD 是一个 64 位的数值类型，具体格式如图 3- 1 所示[1]。使用 VTD- XML 进行 XML 文档解析时，文档中的每一要素(指令、起止标记、属性、数据和注释等)均使用一个 VTD 进行记录。如图 1 所示，为了区分不同要素，VTD 用它的最高四位来标识令牌的类型，具体令牌类型标识参照文献[1]；另外，VTD 中还记录了令牌嵌套深度、长度和偏移量等信息。VTD- XML 通过遍历这些 VTD 记录来找到对应的 XML 信息，然后根据 VTD 中所记录的信息操作二进制数组，完成对 XML 文件

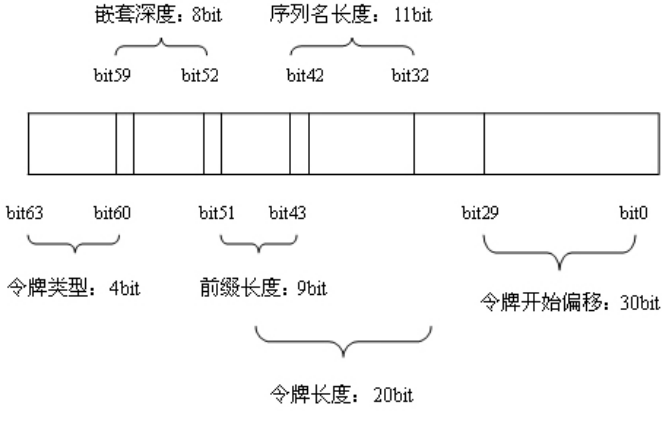

图 1 VTD 的比特层格式

的相关处理。为了正确完成 VTD 记录的遍历工作,它必须借助于位置缓冲 LC。LC 是一个将 VTD 以其深度作为标准构建的树形的表结构。在此基础上,应用程序只需利用相关的信息,便可以在最少的几步内查找到需要的元素,遍历性能十分突出。

表 1 是一组关于 DOM、SAX 和 VTD-XML 之间性能对比的数据[1]。从表中的数据可以看出,VTD-XML 是一种很好的解析技术,不仅速度快,而且占用内存也比较小,适合较大文档的处理。

4 NMR 系统参数处理解决方案对比

为了找到一种高效并适合 NMR 系统的 XML 参数处理解决方案,我们分别使用了以上三种解析方案进行了具体的实现。当遇到较为复杂(如 block 的数量达到了上千个以上)的实验时,相应的 XML 参数文件会达到几十兆甚至几百兆。DOM 解析方案使用的内存一般是原文档的 5~10 倍,经过实验发现它遇到 10 兆以上的情况就没办法处理了,所以它不是一种合适的解析方案。表 2 给出 SAX 方案和 VTD 方案的一组关于 XML 参数文件的测试数据。具体测试环境如下:CPU 为 Intel Pentium 双核 1.73GHZ,内存 2G。

我们发现 SAX 方案与 VTD-XML 方案相比性能有比较大区别,尤其是在处理时间方面。主要的原因是在 SAX 方案中在建立自己的对象模型时伴随这大量的对象创建,而 VTD-XML 是直接操作二进制数组。另外,两种方案当文档达到 100 多兆时均出现了内存溢出的情况,不过就我们项目的实际应用而言,参数大小最多在几十兆,所以两种方案均可胜任。当然,最优方案还是 VTD-XML,它在大文档解析方面优势明显。

表 1 VTD-XML、SAX 和 DOM 性能对比

	VTD-XML	DOM	SAX
处理模式描述	内存中游标是基于非提取的 VTD 记录的	内存中的对象模型是基于节点对象和提取令牌的	低层次的基于提取的令牌
性能	最快	最慢	快,线性速度
内存使用	1.5~2 倍于文档	5~10 倍于文档	不随文档的大小变化
易使用和维护的程度	非常好,对于大部分应用代码短	很好	很差
随机访问	支持	支持	不支持,仅能向前访问
增量更新	支持	不支持	不支持
避免每次都解析文档	支持	不支持	不支持
硬件支持	支持	不支持	不支持

表 2 SAX 方案和 VTD-XML 方案数据对比

层数	文件大小	SAX		VTD-XML	
		耗时	耗存	耗时	耗存
1	55	172	0.8785	134	0.4122
10	169	234	0.5237	156	0.3254
100	1303	640	3.4699	188	3.0328
500	6348	2375	16.401	390	14.2685
1000	12654	4484	30.711	609	28.5660
2000	25266	10953	62.919	1031	56.9088
3000	37678	17524	100.671	1643	94.4551
4000	50491	24096	138.423	2354	130.8429
5000	63103	33734	198.792	3097	189.6715
6000	75715	43584	278.309	3784	267.8157
8000	100940	/	Out of memory	5503	291.6453
10000	162164	/	Out of memory	/	Out of memory

5 结束语

随着 XML 被越来越广泛的应用,合理使用并高效解析 XML 变得越来越重要,尤其对于那些需要处理大量数据的应用程序,显得更加重要。选择合适的 XML 解析技术对系统的系能有着较大的影响,VTD-XML 作为一种新的快速解析技术,必将凭借其性能上的优势被广泛使用,同时,它也将伴随着 XML 技术得到不断地发展和完善。

参考文献:

[1] 刘芳,肖铁军.XML 应用的基石:XML 解析技术[J]. 计算机工程与设计,2005,26(10):308-308.
[2] 鱼雷.VTD-XML.解析技术研究[D].西安:西安电子科技大学,2007.
[3] 沈杰.数字化谱仪软件的研制和应用[D].上海:华东师范大学,2006.
[4] 张亚峰.XML 技术在基于 web 的产品数据管理中的应用[D].西安:西安电子科技大学,2006.
[5] 王惠.基于 SAX 的文档解析技术的研究与实现.开发研究与设计实现[J].2006:147-148.

(上接第 1326 页)

7) 关闭自动更新:右键单击“我的电脑”,点击属性,点击“自动更新”,在“通知设置”一栏选择“关闭自动更新。我将手动更新计算机”一项。

8) 减少开机磁盘扫描等待时间,开始→运行,键入“chkntfs\t:0”

然后连接到 ms 站点顺便升级一次就算优化基本完成,对于 XP 而言,可以采用许多内部命令来看看优化情况,比如 tasklist.exe \svc 可以查看系统服务实际使用情况。

优化一个系统,挺麻烦的。所以我们把它保存起来,我们用 Ghost 生成.GHO 文件,在 Ghost 之前先要作一个事情,清除系统硬件、注册等信息,否则克隆到不同的机子上将无法启动,在 Winxp 安装盘上找 Deploy.cab 中的 sysprep.exe 文件。

执行 sysprep.exe,选择“重新封装”,下面的标记中可以选择“已提前激活”,还可以选择封装完成后是关机还是重新启动。封装完成后,我们再用带有 Ghost 的系统盘启动,用 Ghost 来生成备份.GHO 镜像,备份完成!

在执行封装后,重新开机。

5 结束语

该文从优化 WINXP 的瘦身计划和加速计划以及终止不常用的系统服务几个方面做了详细的阐述。