Java技术的出现和发展促进了Web的迅速普及与壮大，同时，Web的迅猛扩张也反过来催生了与Java相关的其他技术。XML与生俱来的可扩展、跨平台、开放等特性无疑与Java相呼应，形成了完美搭档。在Web这样一个公共的、开放的资源平台与计算环境上，Java技术提供了丰富的实现机制；XML为信息的有效管理和数据集成提供了强大的功能，它提供了一种人和程序都能阅读的描述机制；

基于XML的内容管理

--Web环境下基于XML的信息处理技术

清华大学－IT Frontier株式会社知识工程联合实验室

徐 鹏　李涓子

Web为信息生产者提供了一种便捷、廉价的电子文档发布方式，因此它得以迅速发展，并成为实现全球信息传递与共享的日益重要和最具有潜力的资源平台。而XML的出现使得文本数据的表达有了可遵循的标准，它因此而被称为“Web上的ASCII码”。XML的广泛应用将Web环境下的信息处理技术带入了一个崭新的阶段。

XML与内容管理

在Web产生后的短短几年时间内，传统业务模式发生了重大的变革。这就对信息供应者提出了迫切的要求，他们应当充分利用这种信息交换平台保证其所提供信息的时效性、正确性和完整性。各种各样的商业化网站和信息服务系统能否为信息消费者及时而有效地提供他们所需要的最新信息成为衡量系统技术水准和决定其生存与发展的关键。

企业级内容管理系统可以被看成是一个企业信息架构的基础，其实现得越完善，为用户提供信息的价值就越高。内容管理系统的本质就是从内容角度将信息以具有一定粒度、可管理的“块”或组件的形式通过数据仓库进行存储和管理，系统可以实时对这些组件进行访问、更新和自由组合。基于 Web 的内容管理建立在广义数据库理解的基础上，它是指在 Web 环境下对复杂的内容信息进行有效的采集、组织与集成，实现方便而准确的信息查询与发布。一个基于 Web 应用环境且设计出色的内容管理系统具有开放的商业机遇，这些机遇依赖于对信息内容的复用和加工，使信息能够以多种不同的方式从内容的组织和数据的样式化形式等方面进行个性化处理，并可以利用数据挖掘技术发现蕴含在其中的知识和规则，从而为决策支持提供服务。从技术上讲，Web 内容管理系统的研究融合了数据仓库技术、数据挖掘技术、WWW 技术、信息检索技术、移动计算技术以及多媒体技术等，是一门综合性很强的新兴研究领域。

XML（eXtensible Markup Language）是针对包含结构化信息的文档而设计的一种标记语言。XML 是元语言中的一种，所谓“元语言”，就是能够帮助不同个人和组织定制自己的标记语言的语言，定制后的标记语言可以用于特定的应用领域中实现信息数据的交换。XML 正在成为数据组织和交换的实施标准，并且大量的 XML 数据已经出现在 Web 上。同时，XML 作为一种基础技术在知识管理、通讯管理和数据传送领域扮演着一种重要的角色。不同应用程序之间的数据交换对于开发分布式系统和提供电子商务和灵活性需求来说非常重要。XML 可以将 Internet 转变为一个基于知识仓库的全球计算平台。最终的环境可以被看成是实

现电子数据交换的强大基础架构。一旦采用 XML 表示文档中的元数据，则可以编写一个相关的 XSL（eXtensible Stylesheet Language）文档用以定义元数据的显示方式。通过将在 XSL 中为不同对象定义的规则应用于 XML 数据上，可以实现根据不同用户的不同需求、不同关系或者不同的显示能力，实现不同的数据视图。

基于 XML 的内容管理系统的一个主要特点就是内容“块”仅仅由数据组成（例如文本、图形、表格等），而针对元数据定义的表示信息单独保存。在递送元数据信息的同时提供样式信息的处理方式意味着通过数据管理系统所管理的信息可以很方便地满足不同目标的需求。基于 XML/XSL 技术实现的内容管理系统可以在文档层实现数据模型层与表示层之间的分离。

面向多领域 XML 标准的制订

XML 的诸多先进性令其在产生后迅速得到发展，备受开发者和最终数据消费者的青睐。XML 中的“扩展”一词指的是定义新的标记及其用途的标准机制。由于这一切均是标准化的，所以我们拥有固定不变的途径来描述这些新标记并同其他 XML 用户交流。

利用 XML 技术，数据规范的定义者在充分全面地考虑数据定义完整性出发，定义完整的 XML 数据标准，以满足当前和未来应用的需求。而软件开发者则不必拘泥于固定的脚本语言、开发和设计工具以及数据传输方式，实现一种标准化的、分级别操作的应用环境，在这个环境中不同的工具类软件可以各显神通，从而最大限度地满足客户的需求。包括 IBM、微软、Sun 在内的诸多国际顶级 IT 企业、著名研究机构和国际标准化组织无不对 XML 技术青睐有加，大有得 XML 者得天下之势。他们纷纷积极参与到基于 XML 的数据标准规范的指定和相关软件研发等工作中，几乎每个专业 XML 标准的制订都有该领域在全球占据技术领导权的企业或权威机构参与。

XML数据标准通常是通过词汇表的形式存在的，XML“词汇表”是对XML数据的描述，是元素及其属性、以及你所指定的文档结构的规范。作为信息交换的媒介，它经常是与人类在某种领域的活动息息相关的。XML词汇表的高效性也正是XML应用成功的关键因素之一。目前，针对不同的应用领域的XML词汇表包括科学词汇表、商务词汇表、计算机领域的XML词汇表以及面向其他应用领域的词汇表。

面向内容和具体应用领域的诸多XML国际标准的制订，实现了软件开发人员一个曾经梦寐以求的目标：无论数据产生者位于何处，任何数据消费者都能够通过某种工具与他们交互，并且这种通信是基于数据的含义，而不是数据偶然的表现形式。

XML 数据处理技术的应用

总的说来的面向内容管理领域的 XML 应用可分为五类：

1.应用系统内部的数据

一个大型应用系统内部可能涉及多个数据源，这些数据源包括文件系统、数

据库系统，他们之间的数据格式复杂且异构，同时系统不同功能模块之间所采用的数据模型也可能存在差异。在数据交换过程中将源数据采用基于 XML 的统一数据模型进行表示，可以有效地解决数据访问统一接口问题。由此提出了在多种不同数据源之间实现基于异构数据模型的数据之间转换的研究课题。笔者所在的研究室研发的 XML Transformer 系统在很大程度上有效地解决了关系数据库与 XML 之间以及基于两种异构模式的 XML 文件之间的数据转换问题。

2.实现通用分布式计算环境

XML 技术的应用将改变传统“客户/服务器”工作模式中将运算负荷集中在固定服务器端的模式，而是将其按需分布在客户端和分布式计算环境下的不同服务器上。Web 服务的产生和发展就是这种应用的成功印证。Web 服务已经从面向传统计算设备的应用领域扩展到了移动计算领域。微软与英国沃达丰在今年 10 月 13 日在瑞士日内瓦举行的世界电信联盟展“TELECOM WORLD 2003”大会上宣布将携手进行手机数据服务开发。微软将针对沃达丰的手机服务，提供基于 XML 的 Web 服务解决方案。通过利用 Web 服务，可以在个人电脑中作为应用程序嵌入信息收发、位置信息及收费等手机才有的功能，从而提供扩大应用软件的机会，促进普及可以在个人电脑和移动环境下无缝运行的解决方案。

在 Web 环境下，基于 XML 的数据处理技术的另一个重要的应用领域就是基于 XML 的半结构化信息处理。Web 与传统的文档管理系统结合在一起构成了一个巨大的、异构且分布式的文档仓库，其中比重最大的数据是半结构化文档。传统数据库的检索查询机制以及统计学分析方法已经远远不能够满足半结构化信息处理的需求。笔者所在的研究室研究并开发的半结构化信息智能处理模型 TIPSI（The Intelligence Processor of Semi-structured Information），其研究目标是将内容与样式混合的半结构化文档作为输入，通过对文档知识和元数据的利用将半结构化文档转换为能够提供良好信息复用性且基于 XML 的多视图表示，从而实现针对该类文档的复杂查询服务、基于中介的信息系统和基于代理的应用服务系统。

3.实现不同软件构件的互操作

越来越多的不同类型的应用软件中所采用的数据文件格式已经或者即将采用 XML 标准来定义。在面向各种应用领域的 XML 标准不断推出的同时，相关专用软件的开发市场也呈现飞速发展的态势。

富士通的美国子公司——富士通软件（Fujitsu Software），推出了基础件模块套件“Interstage Suite”。该套件将企业门户、综合软件工具、数据分析软件以及 XML 检索引擎融合在一起。其中“Interstage XML Search”XML 数据的高速检索功能。不论数据大小以及所在位置，均可进行 XML 数据报告、文献的查询。能够同时进行产生 100 项或者 1000 项结果的检索。“Interstage XBRL Processor”作为构筑和配备基于 XBRL（eXtensible Business Reporting Language）的应用软件的工具包，能够对互连网中的多数软件形式和技术间进行自动交换和抽取金融数据。“Interstage Portal”则使企业能够统一各种各样分散型的系统和服务，并向用户、员工、顾客以及合作伙伴提供统一的操作界面。

XML 数据处理软件不但适用于传统以 PC 和服务器为主的计算平台，而且已经延伸到了移动计算平台。在移动计算技术成熟且普及的日本市场中，这类软件不断推陈出新。KDDI 研究所日前成功使用 Java 语言开发出了面向手机的矢量形式图像显示格式定义语言 SVG 的专用浏览器（图 1）。该浏览器依据 SVG Mobile 规范，配备了 SVG 图像的显示、放大、缩小等功能，应用于地图信息服务，可在 KDDI 的手机上运行，用来显示商店、餐馆的位置信息等。

图 1. KDDI 开发的 SVG 浏览

微软将 Office 2003 的设计完全建立在 XML 技术基础之上的举动更证明了将 XML 数据作为应用软件数据文件的定义标准已经成为大势所趋。Office 2003 能够将不同的应用软件数据以 XML 的形式保存下来，其优点在于“能够将独立的数据合并起来”。如果在公司内部不同的服务器上以不同的格式保存数据，则难以搜索到需要的信息或者与别人共享数据。“如果将数据形式统一为 XML，则可以在不同的应用软件之间实现数据的再利用”，并通过在 XML Web 服务和信息工作者之间建立起桥梁，将能够满足顾客的需求。

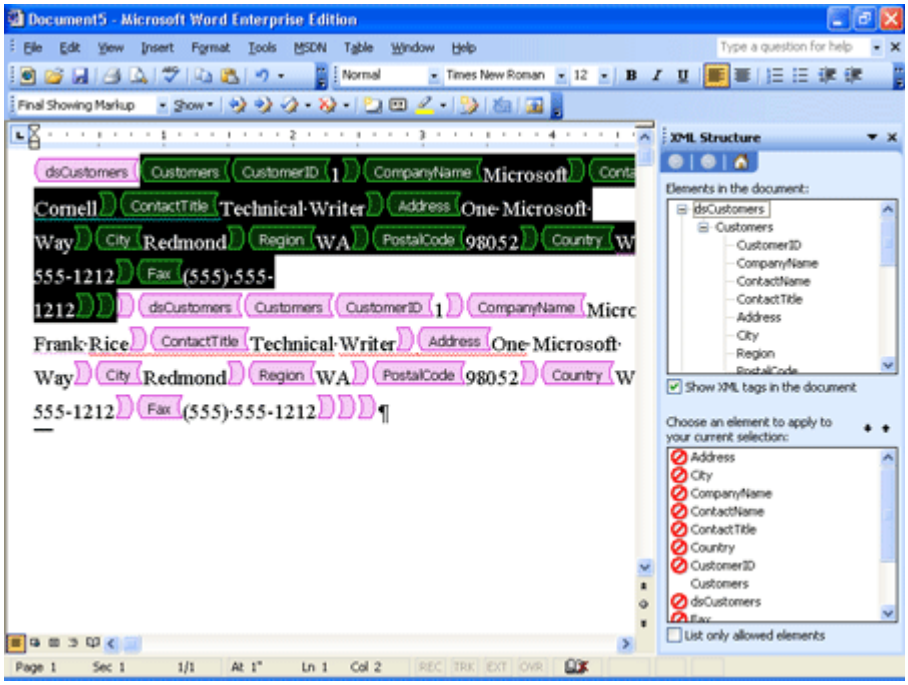

图 2. XML 数据文件的转换

为了开发一个完善的、基于Web的电子商务应用系统，通常需要利用一些消息系统针对具备异构特性的不同后台系统实现和简化系统之间在应用程序和商务处理方面的集成工作。BEA Weblogic Application Server（WAS）作为一个基于XML技术的大型应用服务软件，其中的数据集成套件（Data Integration Suite）以XML为标准格式，实现对不同企业级应用所提供数据信息的集成。

4.可定制发布和数据表示

在基于 XML 技术的数据表示引擎实现之前，如果希望在 Web 环境下生成具有分页特征的文档，则必须使用一种传统的文本编辑软件（例如 Microsoft Word）

来进行文档的设计。而随着 XML 技术在信息系统中的推广，我们可以使用 XML 文件保存纯结构化的源数据，并使用 XSLT 来定义针对 XML 元数据的查询条件和排版信息。通过 XSL 转换处理，即可以得到同时提供所需元数据和样式信息的 XSL-FO 文件。

XSL-FO 规范提供的语法定义了包括字体大小、边缘、间隔、颜色在内的一系列风格设置及一些与页面布局和分页有关设置，它提供了一种比 HTML+CSS（甚至 CSS2）更加完善的可视化布局定义方式。通常，CSS 主要被用于满足 Web 环境下数据简单浏览的需要，而 XSL 格式化对象的设计则适合更加常规的用途。采用 XSL-FO 作为数据表示引擎中定义数据样式化信息的元语言，目标是保证样式化信息能够满足各种用法和表示形式的要求。针对元数据文档实现数据可视化表示处理过程的应用软件称为数据表示引擎（Data Rendering Engine），以下简称“引擎”。

笔者所在的研究室研制的 X2P Server 就是基于 XML/XSL 技术实现的一个数据表示引擎软件，它完成了将 XSL-FO 文件向最终用户可以理解和浏览的形式转换的操作。为了使数据表示过程能够同时适用于不同类型浏览设备的要求，特别是满足种类繁多的移动计算设备的要求，输出格式的可扩展性成为该系统设计所需考虑的首要因素。X2P Server 系统在设计中提出的面向 XSL-FO 的数据表示算法和编译/解释双模式输出过程的设计使系统使引擎能够适应不同类型的 PC 或移动计算设备（例如掌上电脑、支持 Java 的移动电话）的需求。

5. XML 存储技术。

目前，在保存 XML 文档中所使用的主要方法包括支持 XML 的关系数据库、本地 XML 数据存贮器（NXD）、信息内容管理解决方案以及文件系统等。开发人员正在利用 XML 的灵活性和扩展性实现信息内容的共享、再发行、再整合。

美国权威调查机构 ZapThink 的调查结果表明， XML 数据存储技术市场的规模将从 2000 年的区区 7500 万美元，扩大到 2005 年的 41 亿美元以上，上升速度可谓空前。在 2000 年，支持 XML 的数据库管理系统（RDBMS）供应商在市场中所占的比率只有 15％。到 2005 年，这一比率将会达到 65％以上。 2005 年，NXD 供应商的营业额将达到大约 16 亿美元的规模。 NXD 将逐渐成为保存面向文档的 XML 信息内容解决方案的一个选项。还将被用来保存 Web 服务以及 B2B（企业之间的商务）信息等的交易格式。 为了向用户提供利用现有的存贮资产的方法，RDBMS 供应商将继续支持 XML。

基于以上 XML 的典型应用而构造的 Web 应用服务器将能够提供面向半结构化和结构化数据集成处理的完整解决方案。

在“数据共享、信息共享、知识共享”的思想指导下产生的 XML 技术，其产生和发展正走着与 Linux 相似的开放的发展道路。国外 XML 技术正发展得如火如荼，而我们国内在这方面的研究和应用仅仅处于起步阶段。从一些重要的政府机关到包括大专院校在内的研究结构，都在围绕 XML 技术的普及、应用和推广展开工作。同时，国内一些软件爱好者还自发组织了包括 XML 中国联盟在内的网上虚拟技术组织。清华大学计算机系知识工程研究室也在积极与外交部、商务部、中国石化总公司和新华社等单位积极合作，围绕 XML 技术展开高层次的研究和开

发。从全球发展态势看，XML 技术虽然发展速度迅猛，但是其应用仍然处于初级阶段，这对于中国软件业来说无疑是一个迎头赶上的机遇。抓住机遇是我们的责任。