基于 Web 服务和信息检索技术的信息整合方案及其应用*

张晓东
(清华大学 软件学院, 北京 100084)

摘　要: 从国内企业信息管理存在的问题入手,针对数据分散,信息利用效率低下等问题,提出了基于 XML 网络服务和 Office 信息检索技术的一个解决方案,并结合微软员工信息系统(MEIS) 的系统设计实例,介绍了如何利用网络服务和信息检索解决实际问题。

关键词: 信息检索; XML Web Service; XML 架构; SOAP
中图法分类号: TP391. 3　　**文献标识码:** A　　**文章编号:** 1001-3695(2006) 03-0189-04

Project and Application Based on Web Service and Information Search Technology for Information Integration

ZHANG Xiao-dong
(School of Software, Tsinghua University, Beijing 100084, *China)*

Abstract: This article introduces the problems of information management in enterprises, such as information decentralization and inefficiency of data utilization. Then put forward an architecture framework based on XML Web Service and Information Search technology. Furthermore, introduces an example of Microsoft Employee Information System to illustrate how to use Web Service and Information Search technology solve practically problems.

Key words: Information Search; XML Web Service; XML Schema; SOAP

在经济全球化、全球信息化的大趋势下,企业将不可抗拒地加速进入信息网络时代。在现代企业中,由于当前信息来源的多样化,企业中存在大量的异构数据。而传统的各种应用都是直接建立在信息存储层之上的,各种应用直接操作底层的各类数据源,这导致了应用程序不能有效地利用和管理企业的各种异构信息。如果有统一的平台能够整合现有的异构系统,把大量分散、格式不统一的数据进行整合,并为上层应用提供统一的接口和访问规范,那么将会消除这种异构信息所带来的不足。XML 网络服务便是一种能够满足以上要求的标准化技术框架,利用它可以实现企业内部,甚至是企业之间的异构信息和数据整合,实现信息的高效流通和统一管理。下面首先分析一下目前企业信息应用的现状,然后介绍基于 Web 服务的解决方案。

1　企业中信息和数据应用的现状

1.1　企业中的信息和数据

近年来,随着企业信息化的全面发展,企业中拥有大量的数据和信息。数据是企业的宝贵资源,企业经过多年经营已经积累了大量的图纸、文档资料和经验,企业数据资源来自内部数据和外部信息资料,包括诸多方面,如表 1[1] 所示。

如何高效地利用和整合这些数据资源,使企业的数据资源得到再利用,是非常紧迫和重要的问题。企业数据资源建设成为解决这些问题的唯一途径,并由此实现企业数据资源的共享和信息交流。因此,企业大力发展信息系统、办公系统等,希望高效地利用企业中的各种宝贵数据资源,这仅仅是数据利用的第一步。对知识工作者来说,他们需要有效地利用数据和信息,并能够快速转换为相关的个人知识或者企业发展所需要的动力,这样才能发挥这些信息和数据的最大威力。

表 1　企业数据资源来源

企业数据资源											
内部数据							外部数据				
产品图纸	技术文档	技术手册	生产营销	人财物	财务图表	人的经验	相关技术	行业信息	供应商	客户信息	政策法规

1.2　企业中信息处理存在的问题

在目前许多企业中,对于信息和数据资源的建设和利用,主要有以下的一些问题。

(1) 信息和数据的孤岛问题

目前企业中存在着大量信息,但是它们大多处于分散状态,格式不统一,无法有效利用,这是一种信息资源的浪费。很多企业都在利用计算机技术来强化自己的信息管理能力,建立了许多系统和数据库,如财务软件来实现财务信息化管理,用人力资源软件来管理自己的员工等。

这些管理软件为企业的管理带来了很多方便,使系统内部的信息丰富并能够有效利用。但是由于系统与系统之间信息格式不统一,使其独立有余,协同不足,各种信息无法在企业内部或企业之间有效传递或协同利用。信息无法有效地在系统

收稿日期: 2005-03-22; 修返日期: 2005-05-05
基金项目: 国家“863”计划资助项目(2004AA413120)

之间自由流动,形成了一个个信息孤岛,使信息通信的效率低下。现实中,这种信息孤岛既存在于企业内部,也存在于不同的企业之间。

还有一种类型的信息孤岛,其主体不是某个企业系统或者企业本身,而是人员及其相关的信息源。员工是企业竞争力的核心,企业的信息资源中大量有价值的信息来自于员工自身。这些信息有 80% 是存储在员工的头脑、个人设备或者文件中(包括印刷板和电子版),而目前的协同办公系统很难做到真正共享这些资源。

(2) 数据存储介质和信息系统之间存在鸿沟

针对电子的数据和信息而言,目前企业中的电子数据存储介质主要是基于 Word、Excel、文本文件等这样的非规范化档案文件为主。企业的业务是各种各样的,所以通常只能用半结构化或者非结构化的数据去描述,这些文件蕴涵了企业运作过程中重要的文档和数据。

而目前信息系统要求高度规范数据格式,对于输入源的信息和数据要求比较规范的格式。这大大限制了其使用范围,影响了信息高效共享的效率。对于非规范数据和信息,信息系统更多的是采用剔除或者等价变换的处理方式,这大大增加了数据处理的代价,并降低了数据和信息处理效率和准确性。数据和信息的再利用效率低,信息系统整合难度大,为企业的信息处理和管理造成了很大麻烦。

(3) 用户界面操作体验以及可视化的问题

目前,随着企业信息化的开展,开发和部署了一些新的管理系统,用于整合信息,提高信息使用和管理的效率,但是随之带来了新的问题。为了适应新的系统,用户需要改变他们所熟悉的工作流程和方式。而改变用户的工作习惯代价是巨大的,这其中的培训成本,以及由此带来的员工抵触问题都不容忽视,为新系统的应用价值打了一个折扣。如果能够提供一种不改变用户使用习惯,而能够改进信息和数据处理效率的系统,将会大大提高用户的满意度和使用效率。

此外,目前多数的信息系统都是基于数字和字符的表现方式,这样的信息使用难度大,不利于用户快速理解和检索需要的信息[3]。特别是对于知识工作者而言,他们需要的不仅仅是大量杂乱的数据与信息,而是希望能把它们快速转换成目标知识,最好能够直接利用现有的数据和信息自动转换成特定的表现方式,如图表、表格等。这就涉及到数据可视化的问题,这也是现在企业信息系统中还做得不够完善的地方。

(4) 缺乏统一的管理平台

企业中现有的大量数据和信息管理系统之间不能良好通信和整合,缺乏一个统一的管理平台来管理和整合所有这些信息系统。对用户而言,目前信息系统繁多,没有一个统一的访问和使用方式,不能方便地获取需要的信息,而且信息的处理也不能在一个统一的环境中进行。如果能够提供一个统一的管理和使用的接口,使对底层数据和信息的操作对上层应用来说是透明的,那么将会大大提高企业对信息整合和利用能力。

2 XML 网络服务和 Office 信息检索服务

随着网络技术的发展,网络技术慢慢地走向融合和统一。目前,XML 技术成为网络信息通信的工业标准,它也是本文的技术基础。

2.1 XML 和网络服务概述

XML(eXtensible Markup Language,可扩展标记语言)是 Web 上表示结构化信息的一种标准文本格式。XML 语言有两大优势,即自由和超越于格式之上。

XML Web Service 是在 Internet 上进行分布式计算的基本构造块。开放的标准以及对用户和应用程序之间的通信和协作的关注产生了这样一种环境。在这种环境下,XML 网络服务成为应用程序集成的平台。应用程序是通过使用一个或多个不同来源的 XML 网络服务构造而成的,这些服务相互协同工作,而不管它们位于何处或者如何实现[2]。

XML 网络服务的接口可以非常详细地定义,这使用户能够创建客户端应用程序与它们进行通信。这种接口说明通常包含在称为 Web 服务说明语言(Web Services Description Language, WSDL)文档的 XML 文档中。WSDL 文件用于说明消息格式的表示法,以 XML 架构标准为基础,它与编程语言无关。而且以标准为基础,适用于描述从不同平台、以不同编程语言访问的 XML 网络服务接口[4]。

通用发现、说明和集成(UDDI)是 Web 服务的黄页。如果企业自己开发的网络服务希望扩展市场,则需要配置 UDDI 以便能被客户发现。如果 XML 网络服务已经注册过,那么可以利用通用发现、说明和集成(UDDI)来查找,以便潜在用户能够轻易地找到这些服务[5]。

XML 网络服务体系结构有两大优点:①跨平台。允许在不同平台上,以不同语言编写的各种程序和基于标准的方式相互通信。②标准化。使用标准的通信协议,即 XML, HTTP 和 TCP/IP 等,XML 网络服务已经成为一种工业标准。

2.2 Office 2003 信息检索服务概述

Office 2003 中包含了一项非常有吸引力的技术——信息检索服务。从根本上说信息检索服务的后端就是一种网络服务;信息检索服务的前端是 Office 2003 中的所有组件程序(Word, Excel, Outlook 等,微软称为智能客户端)以及 IE 等浏览器。后端的网络服务和前端的 Office 组件程序用 XML 格式的信息相联系,构成了信息检索服务的框架(图 1),展示了一个信息检索结果的界面示例。

这个界面中包含丰富的图形和按钮等元素,它是在 Word 程序右侧的操作窗格中所截取的。Office 2003 的信息检索服务能够提供丰富的界面定制功能,客户端显示的丰富界面都是通过后台的 XML 架构定义的,不需要开发具体的应用程序代码,使它具有与 IE 网络浏览器等瘦客户端相媲美的优点。

Office 2003 预定义了几种 XML 数据架构,用于限定信息流在客户端与 Web 服务之间的传输和数据分析。目前定义的 XML Schema 主要由以下几种[6]:

(1) Microsoft. Search. Registration. request ;
(2) Microsoft. Search. Registration. response ;
(3) Microsoft. Search. Query ;
(4) Microsoft. Search. Response。

它们都是在 urn: Microsoft. Search 命名空间之下,用于完成信息检索服务的注册、查询请求和查询响应等功能所定义的。每一个命名空间下都定义了一些标准成员和框架结构,可根据需要自由扩充自定义类型。这种灵活和统一的定义格式使系统设

计和实现具有更多的选择, 实现了业务逻辑层与显示层分离。

图 1　信息检索服务界面图

但是, 信息检索服务区别于 IE 等瘦客户端最重要的一点是: 信息检索服务能够利用 Office 组件内置的功能和 VBA 编程接口, 实现普通瘦客户端不能实现的程序互操作以及访问系统资源的功能。例如, 可以使用 VBA 接口编写自动生成 Word 图表、Excel 数据统计等功能, 经过安全验证后访问系统中的文件和数据等。这是 Office 信息检索技术强于 IE 浏览器等瘦客户端的最重要一点, 它提供了更强大的功能和灵活性。

此外, 信息检索服务还能让用户在熟悉的办公环境中工作, 绝大多数知识工作者都是在 Office 等应用程序的环境下进行工作, 基于 Office 程序的信息系统是用户最习惯的操作方式, 省掉了培训和系统迁移的成本, 不需要让用户离开他们所熟悉的工作流程和环境, 有效地保证了新系统的价值。

3　基于网络服务和信息检索的信息整合方案

基于以上所介绍的 XML 网络服务以及 Office 2003 的信息检索等技术, 本文为企业信息整合提供一个可行的解决方案, 用于解决本文第 1 节提到的在企业信息管理中存在的几个问题。系统结构如图 2 所示。

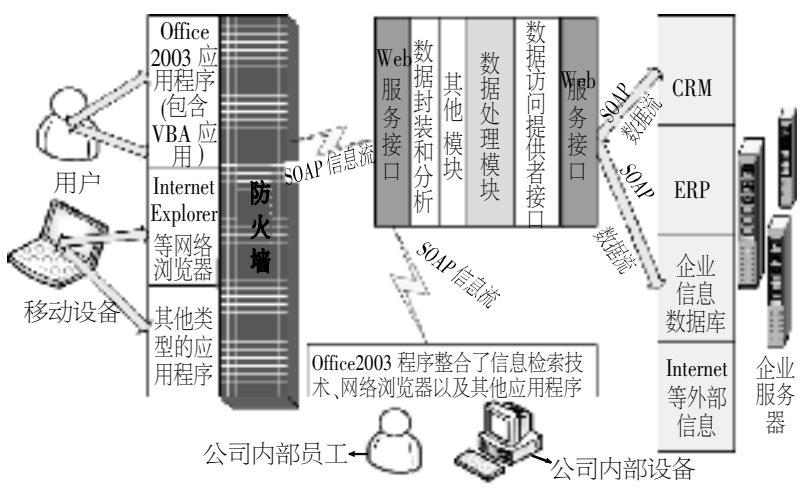

图 2　通用企业信息整合体系结构

这是一种典型的三层系统结构, 数据层、表示层和业务逻辑层互相独立, 每一层都能完成独立的功能, 并为上层提供统一的接口。其中数据层是数据源的汇总, 主要由企业现存的数据库, ERP、CRM、MRP 等系统, 以及企业内部和外部可访问的数据资源组成。这些独立的系统之间无法自由地相互通信[7], 但是它们都可以利用业务逻辑层的数据访问接口模块整合到信息整合系统中, 主要由 Web 服务的相应模块完成。通过一种企业内通用的自定义 XML Schema 数据格式, Web 服务把访问到的数据通过 XML 的数据格式进行封装, 再通过 SOAP 数据流发送到上层应用或者其他的企业管理系统。这意味着企业现存的各种系统可以通过 XML Web 服务进行数据传输和通信, 不会影响系统内部原有的数据格式。

业务逻辑层的功能由 Web 服务完成, 通过 Web 服务中整合的商业逻辑组件程序, 并利用统一的 XML 数据架构模型, 整合各种企业数据系统和信息资源, 把所有的信息和数据整合到统一格式, 使不同系统之间可以通过统一的 XML 数据格式进行通信。这一层主要完成以下功能: ①提供一个统一的访问数据层的数据访问接口, 供其他模块使用, 使数据访问和格式处理的功能完全封装在该模块中, 数据的访问对上层应用来说是透明的。②数据处理模块用于完成数据相关的处理功能, 如检索、分析、知识发现等。这一个模块是信息管理的核心, 在设计中应该考虑可扩展性, 不同的处理算法和流程尽量放置在程序外部, 由配置文件或者系统管理程序管理, 这样数据和信息处理模块的核心功能能够自由扩展。③定义面向上层应用的 Web 服务接口, 为表示层应用程序提供服务, 并为底层的多个信息管理系统提供统一的数据接口, 供现有信息系统使用。④数据格式封装和解析, 该模块主要完成接收和发送的数据进行封装和解析的功能, 具体来说就是使数据能够符合预定义的 XML 架构, 符合上层应用程序定义的统一数据格式。例如, 为了与 Office 通信应该把数据封装成符合信息检索服务的 XML Schema 定义的数据格式。⑤其他模块, 可能包括一些安全加密、系统管理、系统记录等模块完成一些辅助功能。

在图 2 所示的系统结构图中, 最上层为表示层。在传统的 B/S 三层结构中, 表示层由 IE 等浏览器组成, 本文所介绍的表示层将不仅仅局限于传统的浏览器, 它还可以是 Office 组件程序、移动手机、Smart Phone 客户端, 甚至是自主开发的界面应用程序。针对本文而言, 表示层将主要由 Office 组件程序组成, 如 Word, Excel 等 Office 程序。

表示层通过与 Web 服务层的交互, 发送处理命令和数据并接收操作返回的结果。所有的数据传输都是 SOAP 格式的数据封装在 HTTP 数据包中进行的, 数据封装和解析模块用于处理这种预定义的 XML 数据, 并返回能够在 Office 组件程序中显示为丰富界面的 XML 架构数据格式。其中 XML 架构的定义可以根据微软信息检索服务的 SDK 文档作为参考, 为了演示具体的结果, 笔者设计了如下的 XML 架构, 用于定义图 1 的显示界面:

```
<? xml version = "1.0" encoding = "UTF-8" standalone = "no" ? >
  < ResponsePacket xmlns = "urn: Microsoft. Search. Response" >
  < Response domain = "" >
  < QueryId > < / QueryId >
  < Results >
  < Content >
  < Picture Position = ( x, y) . . . >
  < Source > http: / /server/dir/filename. jpg
  < /Source >
  < /Picture >
 . . .
  < /Content >
  < /Results >
  < /Response >
  < /ResponsePacket >
```

以上仅仅是架构定义的基本结构片断, 具体的关于显示标签、图片、下拉列表控件的详细属性的定义这里不再给出, 它们都是在 < Content > 标签中定义的子字段。利用 XML 架构的定义和设计完成数据显示的可视化, 而不仅仅是显示最原始的数据结果。

由于信息检索服务和 Office 整合在一起,所以完全可以利用 Office 内置的 VBA 语言实现程序的互操作性等更加强大的功能。例如,可以把查询到的数据自动生成一个 Word 表格或者 Excel 格式的饼形图等,再把这些数据自动存储在特定的文件中或系统中,只要通过 VBA 接口便可以完成这些任务。还可以通过. NET Framework 的底层框架与 Office 2003 应用程序互操作,实现更复杂的功能,如把信息处理结果自动发送到特定的移动设备或者网络上。

这些高级功能将会满足用户在各种环境下的需求和应用,有效地把灵活性和丰富强大的功能整合在一起。

4 企业信息系统设计实例

上面介绍了企业中信息管理和利用的一些问题,并提出了一个可行的解决方案。下面以一个具体的应用实例来验证这个方案的可行性,并介绍该系统的设计思想。

4.1 背景介绍

在微软大中华区有两千多名员工,原有的员工信息包含在以下的几种系统中:企业活动目录(Active Directory) 、Excel 表格、个人联系人资料、一个基于 Web 的微软全球员工信息查询系统。目前员工的信息获取和管理主要有以下问题:①信息分散,每个系统包含信息的一部分,信息不完整;②信息保存的数据格式不统一,管理和利用效率低下;③用户使用和管理不方便,没有统一的界面和接口,无法实现正确信息的快速获取;④信息更新和维护难度大,数据不同步问题严重。

基于以上所提到的问题,需要设计和开发一套供微软大中华区雇员使用的信息检索系统(Microsoft Employees Information System, MEIS) ,用于解决现有的问题。

4.2 系统设计

根据前面所介绍的基于 XML 网络服务和 Office 的信息检索技术的设计方案,微软员工信息系统(MEIS) 的体系结构如图 3 所示。它从用户视图和组件视图两个角度显示了 MEIS 的外部用户对象、内部系统构成及组成构件间的关联关系。

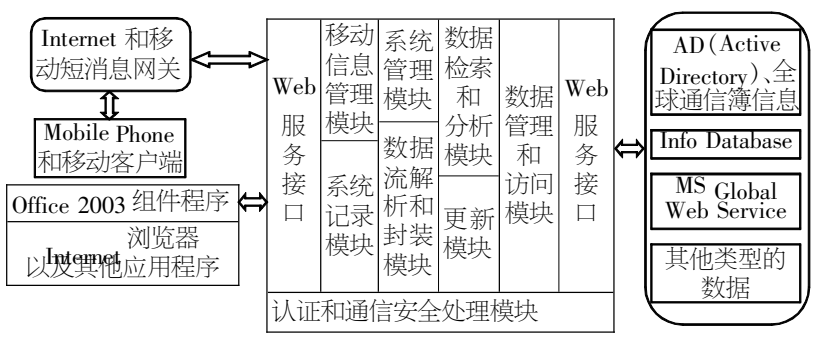

图 3 MEIS 系统结构图

从用户视图来看, MEIS 主要用户有两种,即一般用户和管理员。在系统中也相应设立了两种具有不同权限角色: General User 和 Admin。其中一般用户具有浏览和修改自己信息的权限,管理员用户具有所有员工信息的操作控制权限。

从系统构成角度来看, MEIS 系统是一种典型的三层结构。后端的数据提供者为 AD, MS 全球信息服务和 DB 等。

表示层主要由微软的智能客户端组成,即 Office 组件程序和 Smart Phone 移动客户端,同时提供了传统的 Web 访问方式。

业务逻辑层由 Web 服务构成,主要负责信息的查询、分析、管理、更新等功能,同时提供认证和安全处理、数据封装、数据访问等功能组件。Web 服务采用模块化设计,为功能的扩充和模块的独立化提供支持。

数据访问模块用于连接数据库和其他后台数据资源,提供统一的数据访问接口,同时整合了移动短消息的数据访问功能。

数据分析和检索模块用于从数据库中查询特定的检索词,并分析和返回查询结果给数据解析和封装模块。检索算法和分析算法的核心放在配置文件中,可以通过后台配置文件进行扩充。

数据解析和封装模块主要有两种功能:①把查询结果封装成特定的格式发送给客户端;②把客户端传送来的查询信息,解码成内部信息格式,供数据分析和检索模块使用。

更新模块可以完成普通用户或者管理员提交的更新命令。用户可以修改个人信息,管理员可以添加、删除和更新所有人的信息。

安全管理模块主要完成用户认证和数据加密的功能,并保证信息的安全传输。

移动消息管理模块用于完成移动短消息的发送、接收和信息解析以及编码等功能。除此之外,还有系统管理和系统记录模块完成一些其他任务。

总体来说 MEIS 系统的设计具有以下特点:①分层式结构,典型的三层系统结构,即表示层、业务逻辑层和数据层分离,使系统具有很好的扩展能力。②Web 服务处理过程按照阶段进行分解,进而将问题解决约束在不同的层次,保持各个部分的相对独立性,增强了系统的可扩展能力。③模块化设计,使系统设计具有很好的可扩展性,实现了功能模块的“可插拔”。④统一的标准规范性。在运行平台的设计实现过程中,采用国际相关技术标准规范进行设计和实现,如 SOAP, UDDI, XML 加密和数字签名等;同时在平台内部接口设计中,制定和采用统一的自定义接口,加强了平台总体的规范性。

5 总结

随着网络技术的纵深发展,企业信息网络化是必然趋势,基于 Web 的分布式客户/服务器 Intranet 网络是当今企业处理信息和数据的最佳选择, Web 技术在信息获取、发布上具有不可替代的作用。利用 Web 服务框架结构,并配合其他相关技术(如 Office 技术) ,能帮助企业抛开各类应用系统的对象体系、运行环境、开发语言等技术方面的束缚,打破地域的界限,建立稳定安全的电子信息传递通道,并实现信息和数据有效整合及高效访问和获取。

参考文献:

[1] 吕武. 企业信息化中的数据资源建设方法[J]. 现代管理, 2004, (4) : 101-103.

[2] 代维, 党延忠. XML 技术在办公信息系统中的应用[J]. 计算机应用研究, 2004, 21(1) : 145-147.

[3] 周宁, 文燕平, 等. 信息检索可视化初探[J]. 情报科学, 2004, 22(2) : 138-142.

[4] 杨建武, 陈晓鹏. XML 相关标准综述[J]. 计算机科学, 2002, 29(2) : 25-28.

[5] Roger Wolter. XML Web Service 基础[EB/OL]. http://www.microsoft.com/china/msdn/archives/library/Dnwebsrv/html/webservbasics.asp, 2004-12.

[6] Office 2003 Research Service Software Development Kit [EB/OL]. http://msdn.microsoft.com/office/understanding/research/devdocs/default.aspx, 2004-10.

[7] Vassilis Kapsalis, *et al*. Architecture for Web-based Services Integration[J]. Industrial Electronics Society, 2003, (1) : 866-871.