多语言 Web 网站的结构与实现方法*

杨成甫, 陈　朴, 吴　健, 孙玉芳
(中国科学院 软件研究所 开放系统与中文信息处理中心, 北京 100080)

摘　要: 简化多语言 Web 网站服务的管理与开发。在实践中,管理与开发多语言网站的大部分工作是保持网站的各种信息之间相互独立。在开发与管理多语言网站的过程中有许多与人相关的角色,如设计人员、实施人员(如程序员)、系统管理员、翻译人员与用户等角色。按照这些不同的角色对网站的各种信息进行严格分类,并保持在同一个网站中这些分类后的信息相互独立,也就是说负责翻译的人员不需要看到脚本语言,如 JavaScript。同样,图形设计人员也不需要精通多种语言,也不必在多种语言环境中工作。从以上方面论述如何设计及实现多语言网站。

关健词: 多语言; 本地化; 国际化; 全球化
中图法分类号: TP393　　**文献标识码:** A　　**文章编号:** 1001-3695(2006)02-0131-04

Multilingual Web Architectures and Implementation

YANG Cheng-fu, CHEN Pu, WU Jian, SUN Yu-fang
(Open System & Chinese Information Processing Center, Institute of Software, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The primary aim of this article is to simplify the development and management of a multilingual Web service. In practice, a large part of the development and management of a multilingual Web site consists of attempting to maintain a separation of different kinds of information that constitute a multilingual Web site. During the building and management of a multilingual web site, many human-related roles are involved in, such as designers, builders(programmer), administrators, translators, users and etc. According to these roles, the information of web sites should be classified roughly and saved independently. This means that, translators should not find they have to edit script language, and graphic designers should not have to understand several languages, also not have to work in several languages. In this paper, we will explore the design and implementation of the multilingual web site from the above aspect.

Key words: Multilingual; Localization; Internationalization; Globalization

1　概述

随着互联网络的飞速发展,全球的经济正在走向一体化,企业对用户的定位不再局限在国内,企业在国外的业务比重越来越大。门户网站也逐步成为企业与客户交互的重要途径,用户在网站上享受的服务质量将直接影响企业的服务形象;另一方面,由于中国加入 WTO 以及北京申办奥运会的成功,越来越多的中国企业也逐步开始进入国际市场。因此对这些公司来说,一个国际化的多语言网站将有助于他们拓展国际市场。同时这样的一个网站也为用户提供了一个了解企业的窗口。因此国际化的多语言网站对于他们非常重要。

通常使用不同语言的用户在习惯与文化方面存在着较大的差异,网站既要给使用不同语言的用户提供贴近其文化习惯的界面,又要保持数据与逻辑的统一,以便维护与运营。因此设计与实现一个既能体现不同语言习惯共性又能表达不同语言习惯的个性的模型将十分重要。运用该模型设计出网站的多语言框架,既可以统一处理业务的数据与逻辑,又可以为不同的语言用户提供其独特的界面。

目前大多数网站只提供两种语言,并且对不同的语言只提供分离的网页,没有进行详细的架构设计,而对于这种采取分离的页面的网站,维护与管理工作的费用较大,非常浪费时间,也会导致网站不同语言的页面的信息不一致。

为了有效地降低网站的开发与维护成本,我们从一开始就要考虑多语言,即从网站开始设计时,就要充分考虑在多语言环境中所出现的各种情况,但是这往往是不可能的。下面根据 Web 信息的表现形式,将一个网页中的各种信息分成三大类,对于每一种信息采用不同的处理方法。

在一个 Web 站点中各种信息表现为:格式化信息(如页面样式);网站内容;导航信息(如多语言导航信息)。

在单语言网站中,我们也希望尽可能清晰地将这些不同的信息分开。HTML 标准没有提供任何将内容与格式分开的机制。在多语言网站中,我们还要考虑区域相关和区域无关的内容。表 1 是对这些信息的总结。

通常,我们希望能够将格式化信息、内容与导航信息完全分开,并且在分开的各个部分中还要将区域无关性内容与区域相关性的内容完全分开。然而在实践中,这是不可能完全达到的。对于一个小型的网站来说,建立一个用于保持各种信息分离架构的日常费用是相当大的,所以允许不同的信息相互交叉

收稿日期: 2004-12-28; 修返日期: 2005-03-22
基金项目: 国家"863"计划项目(2003AA1Z2110, 2002AA001033);
中国科学院知识创新工程方向性项目(KGCX2-SW-504)

使用是比较合理。对于一个大型网站来说,管理的费用将会更大,但是可以从这种用来保持各种不同信息分离的方法中得到很多好处。

表 1 信息的说明

格式化信息	具有相似文化与习惯的地区可能使用相同的格式,如所有的西欧地区的人可能使用相同的格式;然而,亚洲可能会需要一个完全不同的格式
内 容	通常情况下,内容是具有很强的地方依赖性;然而情况也不是总是这样,如图形图像可以是全局的。对于一个较为简单的网站,将地方依赖性内容与格式化信息放在一起可能会更好
导航信息	它紧紧依赖于 Web 站点不同的语言版本的并行情况

2 多语言网站的分类

网站的结构与网站内容有着密切的关系。一个对区域有很强依赖性的 Web 网站的各个语言版本,其本地化版本的各个变体之间会完全不同。极端的例子是每个页面的信息只与特定的地区相关。这两种情况下我们可以认为这个网站只是在结构上可能是并行的,而没有做到内容上并行。例如适合一个国家的法律条款不会适合别的国家,按照这样的需求,我们可能做一个结构上并行但内容上不并行的法律网站。

为此,根据网站的并行规模可以对多语言网站进行分类(当然还有别的分类方法):

(1) 对于每个区域都有一个完全独立的站点;

(2) 网站结构并行,但是为每种语言提供一个单独的目录;

(3) 网站结构并行,但是呈现在每个页面上的信息完全不同,每种语言文件通过其文件名表示出来,如 index. en. html 或 index. cn. html;

(4) 网站结构并行,但是有些页面现没有本地化;

(5) 网站结构并行,但是呈现在每个页面上的信息只有细微不同;

(6) 结构与信息完全相同的站点。

下面针对以上各种情况进行详细说明:

情形(1) 不是一个多语言网站,而是一个独立的单语言站点的集合,不同语言的网站保存在不同的服务器,这些服务器可能位于世界的不同的角落,它们之间没有任何联系。

情形(2) 不同语言的页面保存在不同的目录中,所有的页面单独建立。

情形(3) 我们需要将所有的信息单独生成,但浏览与存储结构保持相同。对于网站来说为了有一个相似的存储与导航结构,每个已本地化的页面充当相似的角色,即便是每个本地化页面所描述的信息是不同的。很明显,每个页面必须被标志出它们所充当的角色(或许通过适当的文件名的选择,如中文 index. cn. html; 英文 index. en. html)。

优点: 通过文件名很容易区分不同语言版本的页面,为本地化工作带来很大的方便,翻译人员不用打开每个文件便可知道一个文件用什么语言来显示,如 index. cn. html 文件将用中文将其内容显示在浏览器中。所有的页面内容完全不同,所以没有必要对页面进行翻译,节省翻译方面的费用。

缺点: 做这样的一个网站之前需要用一定的时间对整个网站进行规划,但是仍然会出现页面内容不一致的情形。

情形(4) 这种情况非常普遍,对于一个商业的 Web 网站来说,通常对顶级页面进行了很专业的翻译,但是不对底层页面进行翻译。它们都通过一个缺省的区域页面显示,或是通过机器翻译来显示不同语言的页面。

优点: 这种情况部分解决了网站中不同语言版本页面内容的不一致性问题。

缺点: 通过机器翻译将底层页面进行翻译,由于机器翻译本身的问题,通常底层页面没有被很好地翻译,语种扩展性(一个多语言网站增加新的语言的能力)较差。

情形(5) 信息的单元必须小于一个页面。否则对于特定地区不能表示出什么信息应该显示,什么信息不该显示;页面内容是通过脚本语言从数据库系统中或从文本文件中提取数据来产生,将不同的资源独立保存,提高网站的可管理性与可维护性。

优点: 采用这种方式建设的网站,有很强的扩展性,增加一个语言版本的页面,只要对相应的本地化文件进行翻译就可以完成,可以部分解决网站中不同语言版本页面内容的不一致性问题。相对而言,比前面的几种方案的管理维护费用要少。

缺点: 开发与设计的成本较高,不容易保持网站各种语言内容的一致性。

情形(6) 从一个多语言站点的结构来看这种情形是直接的,但是很少见,又由于没有进行本地化,不适合做一个国际化多语言网站,因此我们就不再讨论这种情形。

在现有的条件下,由于机器翻译的效果不是太理想,所以我们尽可能采用人工翻译的办法解决翻译问题。通过以上分析,情形(4) 采用了机器翻译,节省翻译的费用,比较适合规模较大的网站;而对于小网站来说,页面比较少,人工翻译的成本比较低,情形(5) 则有更强的可操作性,因此本文主要介绍情形(5)。

3 多语言网站的结构

我们希望将找出情形(5) 的不同语言版本之间的共性,同时还要允许不同语言版本之间存在不同的内容,因此情形(5) 是比较复杂的一种情况,同时也是网站比较常用的一种结构。按照情形(5) 做出的网站,该网站的各个语言版本的内容将有所不同。这个网站结构需要的组件如图 1 所示。

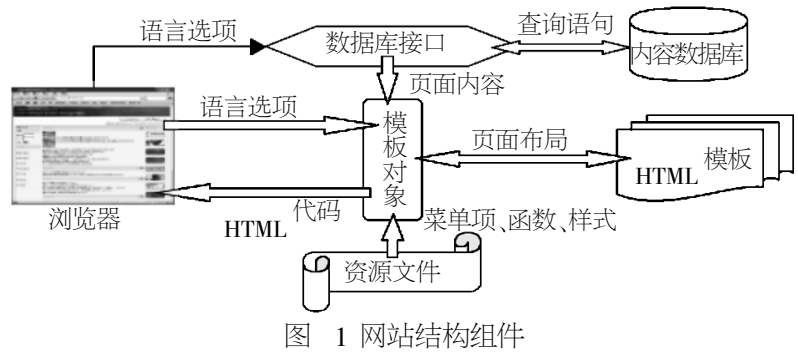

图 1 网站结构组件

首先浏览器向服务器发送请求(HTTP 标准中所描述的格式化信息,包含浏览器的配置信息,如语种选项),服务器就判断是否允许这个浏览器从服务器中检索文档,如果允许就检查浏览器请求的文档是否存在,如果服务器中存在浏览器请求的页面,服务器就将这个页面发送给浏览器。

如果网站没有采用语言协商机制,浏览器的配置信息中的

语种信息将不会被服务器使用, 这样服务器为浏览器提供一个缺省的语言选项, 浏览器用这个语言选项中的语言将页面显示给用户。

数据库接口和模板对象得到语言选项之后, 就分别向数据库、资源文件索取相应语言的资源, 并将这些资源传递给 HTML 模板, 最终生成 HTML 代码, 并将这些代码提交给服务器, 再由服务器将这些代码发送给浏览器, 最终在浏览器显示这些内容。

浏览器(Browser) : 在一个多语言网站的结构中当然要有浏览器, 它是网站与用户之间的接口, 如果没有浏览器, 做这个多语言网站就没有意义。一个浏览器必须能够支持一个多语言网站的所有语言。语言选项是指当前页面所采用语言的种类。例如一个汉语用户, 他希望所看见的是汉语页面, 他的语言选项为 Chinese; 而对于一个英语用户来说, 他的语言项为 English。语言选项是由语言协商机制(Language Negotiation) 或是页面上的与语言无关的独立导航来提供的, 这两种方法还可以结合起来共同为模板对象和数据库接口提供语言选项。

语言协商机制是指浏览器按照用户的浏览器中设置的语言顺序与服务器进行协商, 从而找到一个适当的语言的页面。这个技术优点在于只要用户在他的浏览器中设置常用的语言, 就不需要选择语种便可以自动浏览到他常用的语言的页面。但是, 并不是所有的用户都能够正确地设置浏览器, 通常他们只用浏览器的当前设置, 使我们不能完全依赖语言协商作为语种选择的完整方案。我们就在一个页面上提供一个与语言无关的独立导航。

语言无关的独立导航是要在网页上为用户提供一个语种切换的功能组件——通常是导航条。对于独立导航, 我们希望将它完全与站点中的其他内容分离开, 这样就可以单独管理这个导航条, 就允许我们为不同的语言设置指定不同的策略。

HTML 模板中包含页面布局, 一个 HTML 模板在所有的具有相似格式的语言中共享。资源文件包含与语言相关的资源和库文件, 如菜单项名称、函数库、样式表以及图形图像。制图要将图形的图像层与文本层分离, 这样在重画这幅图的时候只对文本进行翻译, 就可以完成该图的本地化。

模板对象: 它是这个多语言网站结构的核心组件, 负责把从其他组件传递过来的信息组织成 HTML 页面, 提交给服务器, 再由服务器将这个 HTML 页面传递给浏览器。其核心功能是从 HTML 模板中解析出网站页面上的菜单项名称变量, 并与资源文件中的菜单项名称相对应, 从而在页面上显示相应的菜单。

数据库接口: 它负责接收从浏览器传送过来的语言选项, 根据语言选项在数据库查询, 并将结果传递给模板对象。数据库接口可以是一个对象, 也可以是 ODBC, JDBC 等标准接口。

内容数据库存储的是整个网站的内容, 不是完整的 HTML 页面。如果用数据库来保存完整的 HTML 页面, 不仅极不灵活, 而且还要在数据库中保存重复的页面布局, 造成数据冗余。因此, 在我们的多语言网站结构中, 数据库只保存页面的内容, 在显示时通过模板对象将页面内容与包含页面布局的 HTML 模板进行组合, 形成完整的 HTML 文件, 交给浏览器显示。为了保证数据库支持网站用到的所有字符编码, 因此在整个网站中使用 Uincode。

4　实例

多语言网站开发的重要原则是保持网站的各种信息相互分离。针对情形(5) 设计了一个网站(http: / / www. mulitech. org) , 这个网站采用 PHP 作为脚本, 以 PostgreSQL 7. 3 作为数据库, Web 服务器是 Apache。我们将这个网站的各种不同的资源保存在不同的目录中。网站所采用的结构如图 2 所示。

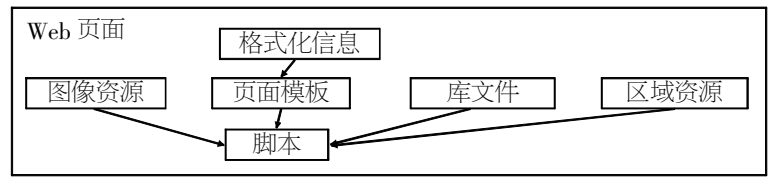

图　2　网站结构图

图 2 中所有的资源文件最终都是要用脚本来处理, 这样在联机的情况下就可以将各种不同的资源组织成 Web 页面。其中:

(1) 格式化信息。使用网页呈现各种不同的效果, 主要包含样式表信息。它可以放在页面模板之中(如果所有的页面可以共享同一种样式) , 也可以与其他资源并行存储。

(2) 图像资源。这里主要是在所有不同的语种页面中都要使用的图形图像, 不包含具有区域相关性的图形图像。

(3) 页面模板。它是 HTML 文件, 主要用于页面布局, 并包含一些变量, 这些变量将被区域资源替代。它的形式如下:

```
< html >
    < head > < title > TITLE} < / title > < / head >
    < body > < P class = " pBlue" > { T_HOME} < / P > < / body >
< / html >
```

其中{ TITLE} 与{ T_HOME} 是两个变量, 它们被包含在大括号中。当用户请求 HTML 页面的时候, 它将被替换成相应的区域值, 如在中文页面中, TITLE 替换成“网站标题”, T_HOME 将被替换成“首页”; 而在英文页面中, TITLE 被替换成“Site tiles”, T_HOME 将被替换成“Home”。由于大括号被用在变量的两边, 起着定界符的作用, 要在网页上显示大括号, 你需要将它替换称为“实体”的特殊字符序列。

(4) 库文件。所有的 HTML 页面都要这些库, 主要有数据库配置信息、SESSION 处理程序、公共函数、常量、构造模板对象程序。

(5) 区域资源。针对不同的语言所提供的一系列文本文件与数据库, 不同语种的文本文件被保存在不同的目录中, 并给于相同的名字。这些文本文件包含与模板中的变量相对应的值, 例如在英文语种中, 文本文件都是 PHP(“PHP: Hypertext Preprocessor”, 超文本预处理器的字母缩写, 它是一种被广泛应用的开放源代码的多用途脚本语言, 可嵌入到 HTML 中, 尤其适合 Web 开发) 脚本, 形式如下:

```
< ? PHP
lang[ 'Home' ] = 'Home' ; / / 与模板中的 T_HOME 变量相对应
lang[ 'Title' ] = 'Site title' ; / / 与模板中的 TITLE 变量相对应
? >
```

而在中文页面中该文本文件的形式如下:

```
< ? PHP
lang[ 'Home' ] = '首页' ; / / 与模板中的 T_HOME 变量相对应
lang[ 'Title' ] = '网站标题' ; / / 与模板中的 TITLE 变量相对应
? >
```

数据库中存放的是页面的内容,如新闻、BBS 中的消息等等,这些资源都是按照语种进行分类存储。

(6) 脚本。用户通过浏览器请求这些脚本,负责将格式化信息、图像资源、页面模板、区域资源和库文件等有机地结合在一起,形成最终的 HTML 页面。

按照上述的方法将各种不同的资源存储在图 3 所示的目录结构中。

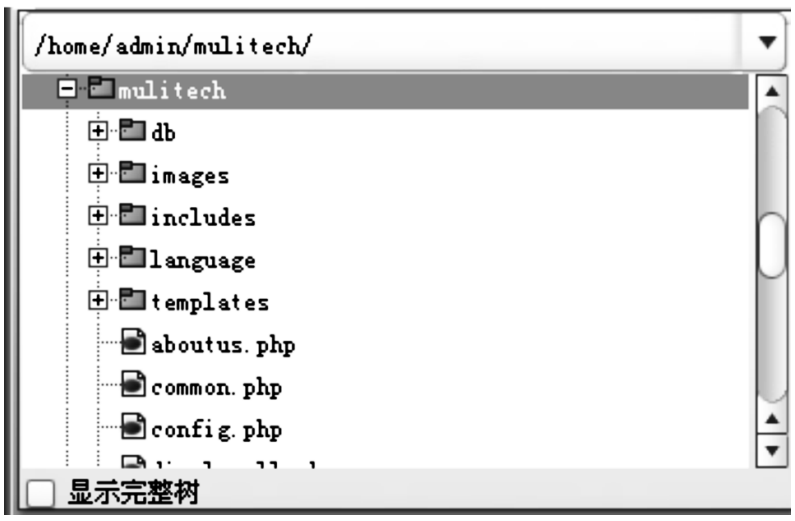

图 3 目录结构

其中,db 目录中保存数据库配置信息、Images 目录中保存图像资源、Imcludes 目录中保存库文件、Language 目录中保存区域资源(主要是语言信息)、Templates 目录中保存页面模板,最后的所有 PHP 文件是脚本。

下面是我们开发的 Mulitech.org 网站的两个截图,图 4 是中文页面,图 5 是藏文页面。

图 4 中文页面

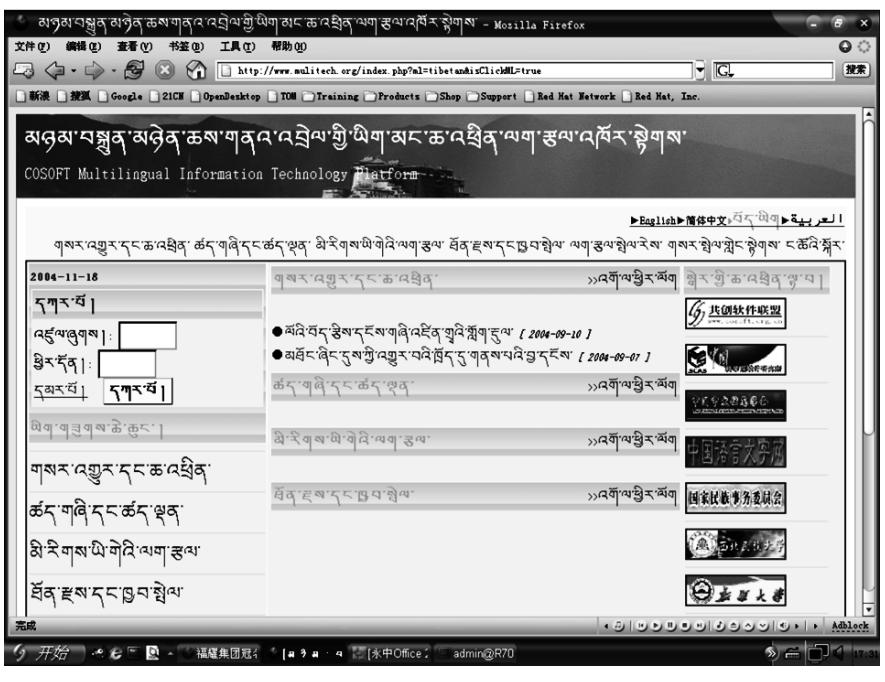

图 5 藏文页面

5 结论与展望

在设计与开发一个多语言 Web 服务的时候,要想节约成本,降低网站的管理与维护费用,最重要的一条是一开始就要考虑多语言,将不同的信息尽可能清晰地区分开,并保存在不同的位置上。其次要考虑你的网站所面对的用户,也就是说你的网站是给什么样的用户提供服务,因为用户的类型(科学工作者或是普通用户)决定了你的网站的类型(是国际化的,还是多语言的,或是国际化多语言的)与风格。所以一开始要尽可多地了解各方面的需求,才能为后来的工作提供一个很好的基础。

这个网站中所有内容的输入都是采用人工的方法,也没有采用机器翻译,必然造成网站的不一致性。目前该网站还没有采用什么方法来解决这个问题,计划是每更新或发布一条新的信息时,给不同语言的管理员发一封邮件通知他们网页更新情况,及时更新网站内容。

目前,对多语言网站的开发还有许多需要解决的问题,如在浏览器上如何显示蒙文,因为蒙文是竖排的,即从上向下的。现在,所有的浏览器,一般只能显示横排的文字(从左向右或自右向左),不支持文本竖排。由 W3C 提出的 SVG(Scalable Vector Graphics)标准很好地解决了文本竖排的问题,但是这只是一个标准,W3C 没有实现这个标准。Adobe 按照 SVG 标准实现了,并以浏览器的插件形式让用户免费下载。

在多语言网站中使用 XML,可以提高网站的开发周期、维护与管理的灵活性。

参考文献:

[1] usan K Miller, Boeing. Defining Localization, Internationalization & Globalization[EB/OL]. http://www.w3.org/International/questions/, 2004-11-30.

[2] W3C. About SVG: 2D Graphics in XML[EB/OL]. http://www.w3.org/Graphics/SVG/About, 2004-10-29.

[3] Jukka Korpela. Multilinguality of a Web Site——The Concept[EB/OL]. http://www.cs.tut.fi/~jkorpela/multi/1-en.htm, 2002-12-21.

[4] Phpbb.com. PhpBB2.0.5[CP]. http://www.phpbb.com, 2004-04-20.

作者简介:

杨成甫(1974-),男,江苏盱眙人,硕士,主要从事多语言 Web 服务方面的工作;陈朴(1979-),男,河南西平人,工程师,硕士,研究方向为民族语言处理;吴健(1962-),男,陕西人,研究员,学士,研究方向为中文信息处理;孙玉芳(1947-),男,江苏张家港人,研究员,博士生导师,硕士,研究方向为操作系统和系统软件、中文信息处理、软件工程。

更正启事

本刊 2005 年第 9 期 76 页上《基于模拟退火的混合遗传算法研究》一文的参考文献[1]中,“周丽,孙树栋”应为“周明,孙树栋”。特此更正。

《计算机应用研究》编辑部

2006 年 2 月