HotJava浏览器汉化方案的设计与实现*

田金兰 徐伟生 吕思飞

(清华大学计算机系 北京 100084)

摘要 HotJava是由Sun Microsystems计算机公司推出的WWW浏览器.基于一种新的面向对象的语言——Java,HotJava浏览器具有与其它同类WWW浏览器无法比拟的动态性能,展示了实现真正的动态网络交互的巨大潜力.本文首先对HotJava浏览器及Java语言作一简单介绍,然后介绍在Solaris平台上开发HotJava中文版的具体方案与实现.

关键词 WWW浏览器,Java语言,HotJava浏览器,汉化.

中图分类号 TP393,TP311

Internet汇集了由多种格式表示的存放在多种主机上的数据,提供了一个巨大的全球信息共享空间.但是,随着Internet的日益膨胀,在Internet上进行信息发现和检索也变得越来越容易迷失而效率低下了.为此,迫切地需要研究在网络上组织和表示信息的更好的方法,以便于信息发现和检索,真正实现交互式地共享网络上的资源和协同工作.1992年WWW(world wide web)的出现成为这方面研究的一个突破.

在WWW上众多的应用程序中,网络浏览器是其核心.网络浏览器集成了数据访问与数据显示的功能,允许用户将分布在Internet上的数据看作一个整体,从而大大方便了Internet上的信息导航.Sun Microsystems计算机公司推出的HotJava浏览器,由于采用了动态的面向对象的Java语言来编写,因而也具有其它WWW图形浏览器无法比拟的动态性能,展示了在Internet上实现真正的动态交互的巨大潜力.

WWW的日益普及使得它必须适应处理多种语言文字的需要.反映在WWW浏览器上,即WWW浏览器必须能够处理包含多种语言文字的文档.为此,自1995-09～1996-04由清华大学计算机系和香港中文大学计算机科学与工程学学系合作开发Solaris 2.4平台上的HotJava中文版,其目的就是在HotJava浏览器中增加中文信息的处理功能,以方便浏览与检索WWW上的中文信息.本文着重介绍这一中文版的设计方案与实现.

1 Java语言及HotJava浏览器

简单地来看,Java语言可看作是C++的一个简化的、安全的、可移植的版本.Java语

---

* 作者田金兰,女,1945年生,副教授,主要研究领域为面向对象语言,系统软件汉化,MIS系统.徐伟生,女,1970年生,博士生,主要研究领域为MIS系统,INTERNET.吕思飞,1968年生,助教,主要研究领域为体系结构,并行处理.

本文通讯联系人:田金兰,北京 100084,清华大学计算机系

本文1996-07-01收到修改稿

言的主要特点可以作如下归纳：

· Java 是一种更简单更易于使用的面向对象的语言.

· Java 编译程序生成与特定硬件软件平台无关的结构独立的字节代码，能够在任何配备了 Java 解释程序和运行系统的处理器上解释或翻译成机器代码执行.

· Java 提供了功能强大的类库来支持与网络协议的接口.

· Java 提供了集成的线程同步功能.

· Java 比 C 或 C++具有更好的动态性能.

Java 语言的这些特点不仅适于开发小的嵌入式 Applet，以增加 WWW 浏览器的交互响应性能和动态浏览能力，它也特别适于在 Internet 这样异构的网络上开发分布式的应用程序.

HotJava 比其它 WWW 浏览器拥有更强的动态性能，它具有把静态的数据转化为动态的应用程序，从而任意地增加新的行为的能力. 与其它浏览器不同的是，HotJava 对不同的数据类型，传输协议和在网络上导航所必须的行为的知识都是抽象的，这些知识松散地组合在一起，带来了更大的动态性和灵活性. HotJava 的动态能力具体表现在以下几个方面：

· 动态的/交互的内容. HotJava 能运行以 Applet 形式包含在 HTML 文档中的可执行内容. Applet 代码既可以是本地的，也可以是分布在 Internet 上的.

· 动态类型. HotJava 的动态行为也体现在它能理解不同类型的对象，并能随时增加对新的对象类型的理解.

· 动态协议. 主机之间用以相互通信的协议是网络上至关重要的一部分. HotJava 只在需要时根据协议名称链入相应的协议处理程序.

· 更适应更新的需要. HotJava 的动态特性使它能在保持网络上的兼容性、互操作性的同时试验新的数据格式和传输协议，而且这些新的成果能在网络上自动、安全地进行安装.

HotJava 之所以具有这些动态性能，是在于它是完全由面向对象的语言——Java 语言编写成的. Java 语言的面向对象特征，以及它的分布性、安全性、动态性、结构独立等特点，使得 HotJava 表现出不凡的动态性能.

2 HotJava 汉化

中文版 HotJava 应具有以下功能：

· 完全支持原有英文版的各种功能. 在处理英文文档时，所提供的功能与英文版 HotJava完全相同；

· 增加中文信息处理功能，包括：

① 输出功能. 能输出包括 GB 和 BIG5 编码在内的任何中文信息. 这些信息可能是 HTML 或其它文档的内容，也可能是 Java Applet 的输出. 至于需要输出中文信息的界面元素，如 Menu，Button，Frame 等，也提供相应的输出功能.

② 打印功能. 能打印含有中文信息(GB 或 BIG5 编码)的文档.

③ 输入功能. 提供对通用输入方法(GB：拼音、内码、五笔字形等；BIG5：仓颉、倚天注音等)的支持，支持 TextField 和 TextArea 内的汉字编辑.

我们在 Solaris 2.4 平台上开发 HotJava 中文版，由于操作系统中缺乏对中文信息处理

的足够支持,所以只能从 HotJava 应用程序本身进行汉化.

针对中文版 HotJava 应具有的处理中文信息的功能,我们应解决的几个问题:

· 中文数据表示的问题. 在汉化一个应用程序时,首先碰到的便是中文字符数据表示的问题. 通常,数据表示的主要问题是字符集和字符集编码的选择. 包括中文在内的亚洲文本含有成千上万个字符,远远超出了单字节字符集(如 Latin-1)所能表示的范围. 所以,必须采用多个字节来表示这些字符. 用多个字节来表示一个字符的两种主要的编码方法是宽字符(Wide characters)和多字节(multibyte). 宽字符编码保持每个字符占有同样多数目的字节,而多字节编码中每个逻辑字符可以有不等长的字节编码. GB 和 BIG5 都是流行的宽字符编码,它们分别对应简化字符和繁体字符. 在这两种编码中,每个逻辑字符(一个汉字)均用两个字节来表示. 因此,在我们的应用程序中,所有的中文数据操作模块均是围绕着宽字符编码方式来构建的.

· 中文字符显示的问题. 关于中文字符的显示,首要的要求就是字体的映射(font mapping). 在 Solaris 2.4 平台上,由于 X11 已经提供了对 16 位字符字体的标准支持,所以在相当大的程序上简化了中文字符输出的工作. 尽管日期、货币单位、重量和度量的习惯表示方法等问题也是发展本地化的语言版本时应注意的与输出相关的方面,但我们在汉化 HotJava时,出于具体的实现方案的考虑,对此未加深究.

· 打印中文字符的功能. 由于 HotJava 在打印文档时能自动生成含有中文字符的标准格式的 postscript,因此若配备一个中文打印机或安插一个小的转换程序以利用已有的英文打印机,即可从 HotJava 中打印出含有中文信息的文档.

· 中文输入和编辑. 在中文 HotJava 中,TextField 和 TextArea 要求能进行中文字符的输入和编辑. 与英文字符相比较,汉字字符数目巨大,汉字的输入必须借助有效的输入方法把用户在标准 QWERTY 键盘上输入的键序列转换成汉字在机器内部的表示形式. 输入方法的实现涉及将键序列与输入表的表项逐项进行匹配以找到正确的内部编码. 另外,汉字输入的实现也必须考虑到动态增删输入方法的可能性. 汉字的编辑主要要针对一个汉字由两个字节表示的特点,消除产生半个汉字的情况. 在 Solaris 2.4 平台上,由于目前的 X11 和 Motif 版本都还缺乏对汉字输入和编辑的有效支持,所以完全要自行实现,难度和复杂度都较大.

在 X Window 上开发支持多种语言编码的应用程序,存在着 2 种不同的途径:国际化(Internationization,简称 I18N)途径和本地化(Localization,简称 L1 ON)途径. 我们采取了本地化 HotJava 的实现方案,用成熟的本地化技术对每种语言编码分别进行处理,由于该 HotJava 中文版需支持 GB 和 BIG5 两种中文编码,在未来也还可能增加新的编码支持(如与中文编码相近的日文和朝文),所以实际上我们的中文版 HotJava 被设计成一个支持多种语言编码的本地化(Multi-Localization)应用程序,程序中提供接口以在不同的语言编码之间进行切换.

HotJava 是一个完全由 Java 这种面向对象的语言书写的应用程序,在 Solaris 平台上,其层次关系可用图 1 表示.

在汉化 HotJava 的过程中,我们的主要工作是修改扩充 Java 类库中与平台相关的部分,因为汉字的输入输出主要涉及的是与平台相关的界面元素. 概括为:

· 在HotJava界面中增加菜单选项Encoding，以选择编码(Latin-1，BG或BIG5)；增加对中文编码的处理功能；

· 修改所有Java界面元素类(如Font，label，Button，Menu，Frame等)中所有有关字体设定、字符串输出，计算字符串宽度等与文档显示布局有关的方法；

· 以cxterm作为一个嵌入的模块来支持各种输入方法；自定义能编辑中文字符的Motif类XxTextField和XxTextArea，这些类与cxterm服务模块通信，共同完成汉字字符的输入和编辑；修改Java的TextField和TextArea类与这些自定义Motif类的接口.

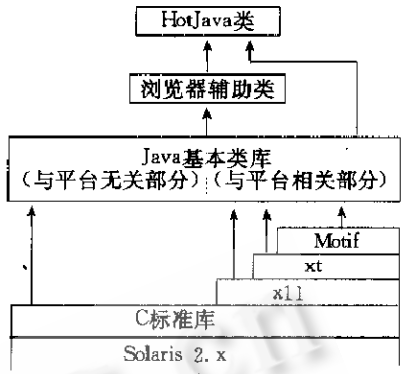

图1

3 结束语

HotJava浏览器由于采用了新的面向网络编程的面向对象的语言——Java，因此具有很好的动态性能. WWW的迅速扩张使得WWW浏览器必须具备处理多种语气文字的能力. 由清华大学和香港中文大学合作开发中文版HotJava是使HotJava适应处理中文信息的需要的重要一步，也是使得它适应WWW的多语言化趋势的重要尝试.

参考文献

1 Sun Microsystems Computer Company. The HotJava Browser: A White Paper, May 1995.
2 Sun Microsystems Computer Company. The Java Language: A White Paper, May 1995.
3 Nicol Gavin T. The multilingual world wide web. Electronic Book Technologies, Japan.

DESIGN AND IMPLEMENTATION OF A CHINESE HOTJAVA BROWSER

TIAN Jinlan XU Weisheng LU Sifei

(*Department of Computer Science Tsinghua University Beijing* 100084)

Abstract HotJava is a WWW browser, which is developed by Sun Microsystems. Based on Java language which is a new object-oriented language for network programming, HotJava browser has dynamic and interactive behavior above and beyond other WWW browsers. In this paper, HotJava browser and Java language are simply introduced, then design and implementation of a Chinese HotJava is described.

Key words WWW browser, Java language, HotJava browser, Chinese.
Class number TP393, TP311