内容全球化管理系统的设计与实现

吴国仕 何所惧

(北京邮电大学 软件学院, 北京 100876)

摘 要: 针对企业多语言内容信息管理的需求,提出了用于检测客户端内容信息变化的数据库内容变化监测算法与文件内容变化监测算法,并结合 XML, CORBA (Common Object Request Broker Architecture) 及工作流技术,设计并实现了能够有效集成现有的企业多语言内容的管理系统. 结合多语言内容信息的特点,完成了系统整体框架、系统主要模块的工作流程及内容变化检测算法的设计;通过系统的实现, 验证了该算法是有效的及系统设计是可行的.

关 键 词: 信息管理; 基于内容的信息提取; XML; 算法

中图分类号: TP 317. 1

文献标识码: A **文 章 编 号**: 1001-5965 (2007) 02-0233-05

Design and implementation of global content management system

Wu Guoshi He Suoju

(School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The algorithms of "change detection for database content" and "change detection for file" were put forward for detecting the change of client content data based on the actuality of multilingual contents in the enterprises. An effectively centralized multilingual content management system was designed and implemented by using the XML, CORBA (common object request broker architecture) and workflow technologies. Based on the features of the multilingual contents, the system framework, workflow processes, content change detection algorithms and system implementation were designed in detail. The feasibility of the algorithms and system design was verified by implementing the system.

Key words: information management; content based retrieval; XML; algorithms

随着经济的全球化, 我国本土的一些正在向国际化迈进的大中型企业逐渐积累了大量的不同种类、不同语言的内容信息. 面对这些庞大的内容信息, 目前企业的 IT部门大多采用人工管理的办法;而随着内容信息的几何级数增长, 人工管理的模式越来越显得力不从心[1-2]. 这主要是由于内容信息的如下特点所引起的: ①内容信息的多样化:企业大量的以非结构化形式散落在企业各个角落的内容信息, 诸如各种报表、账单、电子文档、网站的各种元素、图片、传真、扫描影像,以及大量的多媒体的音频、视频信息等; ②内容信息的多语言化:全球化企业的内容信息是指企业所拥有的内容信息具有多种语言的版本[3]; ③内容存储的分散化:由于企业的内容文档是随着企业的发展而逐渐积累起来的, 因此它所拥有的内容文档可能存储于不同的机器上,或者使用不同的存储方式, 诸如文件系统、数据库系统等;同时企业的不同部门也可能拥有自己的内容文档; ④管理过程的复杂化:由于上述内容信息的特点, 给内容信息的全球化管理带来了很大困难. 特别是多语言内容信息的管理,它涉及到信息的国际化、本地化以及多语言内容的同步化等方面.

内容信息的上述特点, 向企业提出了一个新的挑战,那就是如何在保持全球统一的企业品牌的前提下,有效地开发、利用和管理多语言、多区域及多文化的企业信息;而这种多语言内容管理

收稿日期: 2006-02-23

基金项目: 教育部重点实验室开放基金资助项目 (F0607-22)

作者简介: 吴国仕 (1957 -),男,安徽阜阳人,教授, guoshiwu@yahoo. com.

的好坏,直接影响到企业的国际信誉和市场地位.

当前,国外一些从事多语言服务的 IT公司正与一些翻译服务公司联合,为大公司客户开发全球化信息管理的解决方案,如多语言翻译管理软件龙头企业 TRADOS同欧洲翻译公司 SDL合作开发的 GM(Global Information System), idiom公司的产品 WorldServer,它们都能较好地解决企业的全球化信息管理问题[4-5],但价格昂贵,而国内目前还没有类似的产品或解决方案问世.由于经济全球化的需要,目前一些具有前瞻性的国内企业已开始考虑在现有的企业内容资源的基础上,将企业的多语言信息进行有效的管理以便进一步增强企业的国际竞争力,这就是本文所研究的目的.

1 系统的工作原理及设计

1.1 系统的基本设计思想

当前国内企业的内容管理主要模式包括基于文件、基于数据库及基于传统的 CMS(Content Management System)的存储方式.本系统正是针对企业多语言内容信息不同的存储方式及内容信息分散性的特点,采用了一种多服务器协同的框架,及时地监测原语言内容信息的变化;并结合工作流管理系统,对企业分散的多语言内容信息实行有效的中心化管理和多语言内容信息的同步.

1.2 系统的主框架描述

该系统通过采用 XML, CORBA (Common Object Request Broker Architecture) 及工作流技术,利用主服务器的项目管理、调度及协调功能,数据适配器的内容变化监测及数据提取功能,及工作流管理系统灵活的过程建模功能,将分布于不同的机器上,以不同方式存储的内容信息进行有效的中心化管理,从而实现了多语言信息同步的自动化.该系统的数据适配器端可根据企业内容信息的存储方式,或根据企业的要求进行配置;并可在一台机器上支持一种及几种数据适配器的共存.系统主框架见图 1,主要包括以下 4个部分:

1) Global Content Server(主服务器).它包括 App Server(应用服务器)与 Workflow Server(工作流服务器).

应用服务器主要包括 WEB服务、API、业务逻辑管理、项目管理等.其主要功能是创建用户界面;管理用户创建的翻译项目;协调主服务器与适配器之间的关系;管理内容资源及系统所用的翻译、校对及审批人员资源;根据业务逻辑的要求实现模块调度.

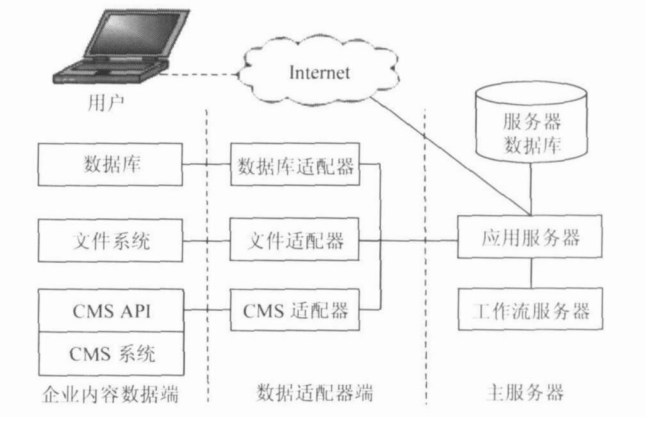

图 1 系统的基本框架

工作流服务器主要负责管理用户创建的工作流过程及工作流过程实体的运行与调度,包括工作流引擎、工作流程编辑器、工作流程监视器、系统管理员及用于客户端调用的 API[6-7].

2) Data Adapter(数据适配器端).它是应用服务器与企业内容信息沟通的桥梁,其主要功能包括:监视企业内容信息的变化、传送数据信息与控制信息到应用服务器、接收应用服务器的指令、完成应用服务器指定的任务.内容适配器包括数据库适配器、文件适配器及 CMS适配器.

3) Enterprise Content Data (企业内容数据端).它是企业所拥有的多语言内容信息的存放处,包括文件系统、数据库系统及传统的 CMS系统;其存储结构应当支持多语言内容信息的存放.

4) Server Database (服务器数据库).它是主服务器的重要组成部分之一,主要用来存储 3部分的数据:①主服务器资源管理、项目管理等所需要的数据;②Workflow Server所使用的数据;③用来记录企业内容信息的敏感数据,其中包括内容信息的 ID、内容信息的文件名、该内容信息递交到主服务器的时间、内容信息的文本内容等有关数据,这些数据供监测企业内容数据变化使用.

1.3 适配器端关键模块流程及算法

1) 适配器端工作流程.适配器的主要功能是定期地启动数据适配器端的工作流程,在流程中通过调用内容变化监测模块确定企业内容信息的变化情况.一旦发现源内容信息发生变化,适配器产生一个供主服务器进行管理和调度使用的 XML格式的项目控制文件和数据文件.随后,数据适配器将项目控制文件及数据文件送主服务器.当适配器是数据库适配器时,送主服务器端的数据文件包含客户端数据库中发生变化的源内容信息;对文件适配器,送主服务器的数据文件是内容发生变化的文件.当主服务器收到请求后,由主

服务器创建翻译项目并启动其自身的工作流程,在工作流程中主服务器将需翻译的数据文件送相关的翻译人员进行翻译处理.然后,主服务器抽取项目控制文件及数据文件中的有关内容,并把适配器端进行数据变化检测所需要的重要信息存储到主服务器数据库,为适配器下一次的数据变化监测做好准备,适配器端的工作流程见图 2.

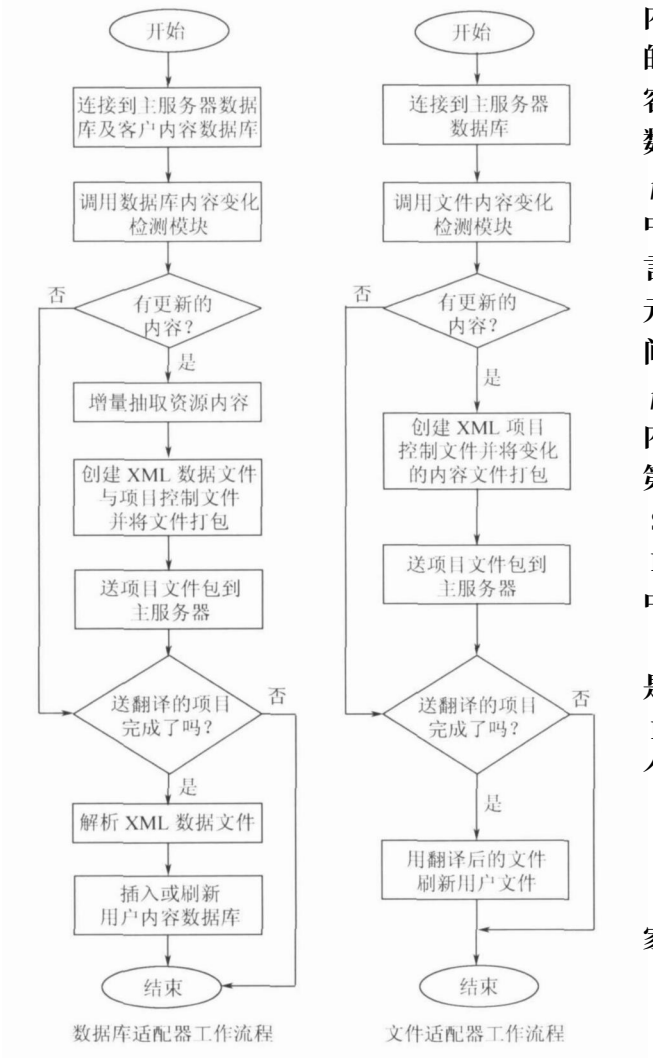

图 2　适配器端工作流程

当适配器完成了内容变化检测及变化的内容信息的抽取后,它开始检查是否有来自主服务器的已完成的翻译项目.

对数据库适配器,一旦发现翻译完成的项目,它首先对主服务器送回来的翻译成其它语言的包含内容信息的 XML 文件进行解析,抽取翻译后的内容信息单元,并用新翻译的内容信息去刷新用户的内容数据库.对文件适配器,一旦发现翻译完成的项目,它将把从主服务器送回来的翻译的内容文件存储到客户的文件系统中.

2) 数据库适配器内容变化检测算法.设 DB_s 是主服务器端的数据库; T_{s_prj}是 DB_s 用于记录企业数据库信息变化状况的表; F_{pid}, F_{scc}, F_{slc}, F_{tcc}, F_{tlc}, F_{tcid}, F_{tc}, F_{ltdt}, F_{tcl}和 F_{tccs}是表 T_{s_prj}的字段.其中, F_{pid}为项目的 ID, F_{scc}为源内容信息国家代码, F_{slc}为源内容信息语言代码, F_{tcc}为目标内容信息国家代码, F_{tlc}为目标内容信息语言代码, F_{tcid}为内容信息单元 ID, F_{tc}为内容信息单元文本串, F_{ltdt}为内容信息单元最后刷新日期时间, F_{tcl}为上次递交的内容信息单元的串长度, F_{tccs}为上次递交的内容信息单元串的 Checksum;设 DB_c 是客户端内容数据库; T_{c_source}是 DB_c 中用于存储内容信息的表; f_{cc}, f_{lc}, f_{tid}, f_{tc}, f_{cdt}, f_{ludt}, f_{st}, f_{tccs}是表 T_{c_source}的字段.其中, f_{cc}为内容信息的国家代码, f_{lc}为内容信息的语言代码, f_{tid}为内容信息单元的 id, f_{tc}为内容信息单元的字符串, f_{cdt}为内容信息单元的创建日期及时间, f_{ludt}为内容信息单元的最后刷新日期及时间, f_{st}为该内容单元是源文本单元, f_{tccs}为最后刷新的内容信息单元的 Checksum;设 $T_c[i]$是 T_{c_source}的第 i个记录; $T_s[j]$是 T_{s_prj}的第 j个记录;

```
Set change_detected = 'false';
For each Tc[i]∈Tc_source do //对客户数据库内容表中每一个数据;
If (Tc[i].fst = 'true') then { //如果该内容单元是源语言内容单元;
For each Ts[j]∈Ts_prj //对主服务器 DBs 表中每一个数据;
    If ((Tc[i].fcc = Ts[j].Fscc) &&
        (Tc[i].flc = Ts[j].Fslc)) then
    { //确认是否是同一种语言编码和同一种国家编码;
    If(
      (Tc[i].ftc. length≠Ts[j].Ftcl) ||
      (Tc[i].fludt > Ts[j].Fltdt) ||
      (Tc[i].ftccs ≠Ts[j].Ftccs))
    then { //判断内容信息串的尺寸、最后变化时间及 Checksum 是否发生变化;
        change_detected = 'true'
        return change_detected;
      }
    }
    End For
    }
    End For
    Return change_detected
```

3) 文件适配器内容变化检测算法.当适配器是文件适配器时,送主服务器的数据文件是一个

或多个需翻译的文件. 客户端源文件及目标文件的存储结构见图 3.

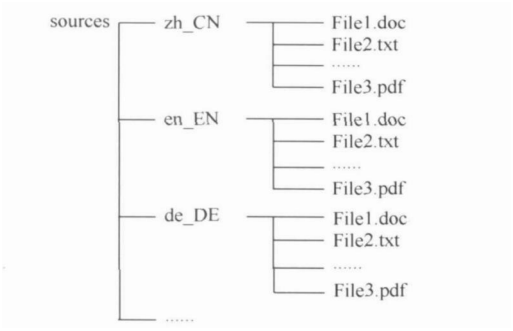

图 3 企业客户端的文件结构示意图

设 T_{fprj}是 DB_s中用于记录客户端文件变化信息的一个表;设 F_{pid}, F_{fcc}, F_{flc}, F_{ftcc}, F_{ftlc}, F_{fid}, F_{fp}, F_{fn}, F_{fludt}, F_{fl}, F_{fcs}是表 T_{fprj}的字段. 其中, F_{pid}为项目的 ID, F_{fcc}为源文件的国家代码, F_{flc}为源文件的语言代码, F_{ftcc}为目标文件的国家代码, F_{ftlc}为目标文件的语言代码, F_{fid}为源文件 ID, F_{fp}为源文件 PATH, F_{fn}为源文件名, F_{fludt}为源文件上次递交的日期时间, F_{fl}为源文件上次递交时的尺寸大小, F_{fcs}为源文件上次递交时的 Checksum;设 P是一个包含源语言文件的集合; V是包含文件属性的集合,且有 V(cc, lc, fn, fp, s, d, cs). 其中, cc为该文件国家代码, lc为该文件语言编码, fn为文件名, fp为文件 PATH, s为文件尺寸大小, d为文件的最后变化时间, cs为文件的 Checksum;设 $T_f[i]$是 T_{fprj}的第 i个记录, $P[j]$表示文件集的第 j个文件, $V[j]$表示对应于文件 $P[j]$的属性.

Match($T_f[i], V[j]$)是一个判断 $T_f[i]$与 $P[j]$所指向的文件是否属于同一个文件的函数,它主要从文件名、文件 PATH及文件编码等方面进行判断.

```
Set change_detected = 'false';
For each P[j]∈P do //对客户端文件系统中的每一个文件;
    Get V[j];
    For each T_f[i]∈T_fprj do
    If( Match(T_f[i], V[j]) ) then { //如果 T_f[i] 与 V[j]所描述的是同一个文件;
        If( (T_f[i]. F_fludt ≠V[j]. d ) then { //如果文件的变化时间与主数据库中纪录的最后变化时间不同;
            If ( (T_f[i]. F_fl ≠V[j]. s) ||
              (T_f[i]. F_fcs ≠V[j]. cs ) )
            Then { //如果文件尺寸大小或 Checksum //与主数据库中记录的不同;
                change_detected = 'true';
                return change_detected;
            }
        }
    }
    End For
End For
return change_detected
```

4) CMS适配器算法. 由于当前世界主流的 CMS产品 Intervowen, Documentum, FileNet等基本上是以文件为基本单元来进行文档内容管理的,因此,本文的 CMS适配器算法基本上与文件系统适配器算法相同,只是在文件属性的检测时需通过调用 CMS的 API来实现,因此,该算法不再赘述.

5) 适配器端设计的其它说明. 在适配器设计时,考虑到主服务器的负荷,把定期进行的内容变化的监测工作放到适配器端;在充分分析了企业内容数据特点及确保内容信息变化不被漏报的基础上,尽可能地简化内容变化的监测算法,以期减少用于内容变化检测的时间开销. 上述算法通过运用 XML模式特点,针对企业内容信息的多样性,利用主服务器端的数据库将企业的内容信息的特征进行集中管理,它具有灵活性强,时间开销少,实用可行等特点.

1.4 主服务器端的工作原理及流程

主服务器的应用服务器的主要功能是创建用户界面、管理用户创建的翻译项目、协调主服务器与适配器之间的关系. 工作流服务器系统是为了增强内容信息全球化管理的灵活性而设计的,它支持 WFMC标准,具有过程版本控制、资源管理、角色管理、权限管理等功能. 除此之外,主服务器提供了方便用户的工作流程再造的 UI界面. 当前主服务器所使用的工作流程见图 4.

该工作流可以在两种情况下启动: 一种是当适配器端发现内容信息变化的情况下通知主服务器,由主服务器调度模块启动主服务器的工作流程;另一种是用户可以使用 Internet Explorer界面,通过主服务器界面手动启动工作流程以便递交新的翻译项目. 下面对该工作流程的主要任务进行简要的说明: ①工作流启动后,主服务器将等待接收适配器端发送过来的内容信息源文件和项目控制文件组成的文件包; ②翻译预处理:预处理按照项目控制文件所提供的信息完成文件格式的处理及按客户需求进行专门的客户化处理,并将内容信息文件中的有关信息存储到主服务器的数

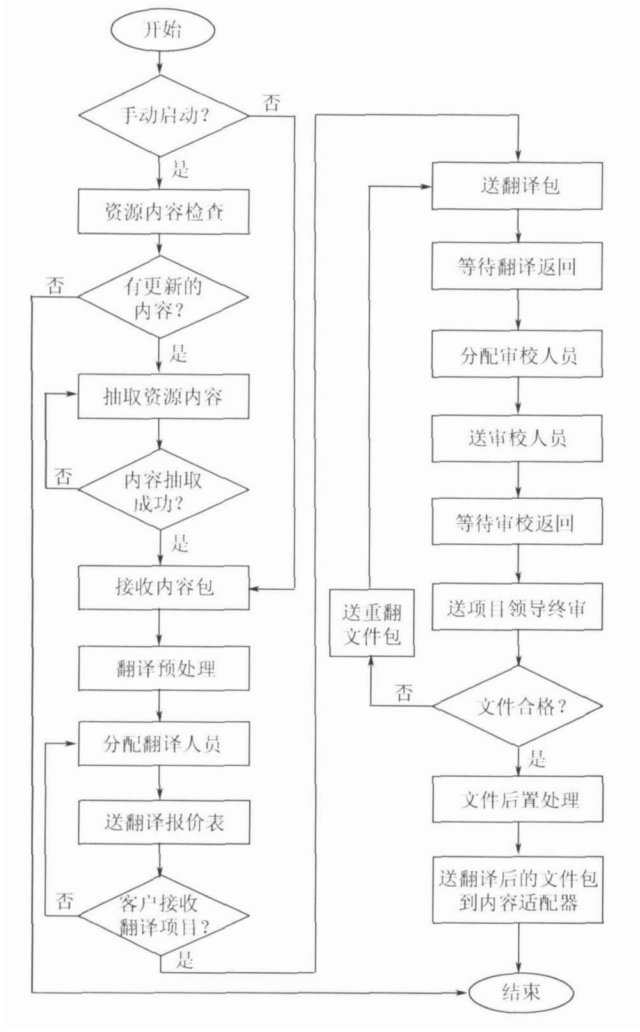

图 4　主服务器工作流程图

据库,以供下次适配器端检测内容变化时使用;③分配翻译人员:从资源库中查到和本翻译项目相匹配的翻译人员,由项目管理员安排其去完成这项任务;④按翻译字数送报价表到翻译人员:如果该翻译人员拒绝该项目,将重新分配新的翻译人员去完成这项任务;⑤文件的后置处理:规范化翻译后的内容文件,登记项目完成时间到主服务器数据库并检查翻译后文件的 XML格式,并重新对翻译后的文件打包;⑥发送翻译后的文件包到内容适配器:将文件包发送到内容适配器,由适配器对此解包,并按控制文件的要求将翻译后的内容替代原有的目标语言的内容信息.

2　系统的实现

考虑到企业内容信息存储在不同的操作系统中,本系统采用了基于 J2EE的开发平台.主服务器与适配器之间采用了 CORBA协议,工作流引擎采用自己开发研制的工作流管理系统,主服务器使用的工作流程如图 4所示,随着企业需求的变化,可通过工作流程编辑器实现流程再造.实现后的系统经过实验室测试,效果良好,能够有效地对存储在数据库及文件系统中的多语言内容信息变化进行有效的监测,并实现变化内容的信息同步.

3　结 束 语

本文针对企业多语言内容信息的特点,提出了数据库内容变化监测及文件内容变化监测两种算法,并结合 XML, CORBA及工作流技术,成功地设计并实现了适用于企业多语言内容全球化管理的系统.该系统具有灵活性强、客户化方便、能较好地集成与管理企业现有的多语言内容信息的特点,并能方便地与企业现有的多语言信息发布平台集成.该系统的使用会进一步增强企业内容管理的自动化程度、减少企业 IT人员在手工管理多语言内容信息方面带来的失误、促进企业内部多语言信息的同步,为提高企业在国际市场上的知名度和竞争力打下良好的基础.

参考文献 (References)

[1] 张婵,罗佳.企业内容管理综述[J].现代计算机, 2005, 27(8): 17 - 19
Zhang Chan, Luo Jia. Summary of the enterprise content management[J]. Modern Computer, 2005, 27(8): 17 - 19 (in Chinese)

[2] 石雪松.内容管理的真正内涵[J].中国计算机用户, 2003, 548(4): 25 - 27
Shi Xuesong. The real means of content management[J]. China Computer Users, 2003, 548(4): 25 - 27 (in Chinese)

[3] Budin G. Global content management-challenges and opportunities for creating and using digital translation resources[EB/OL]. Vienna: University of Vienna, 2002 [2006-02-03]. http://www.ifi.unizh.ch/cl/yuste/postworkshop/repository/gbudin.pdf

[4] Lawlor T. Creating managing and maintaining a global website[EB/OL]. California: SDL International, 2005[2006-02-13]. http://www.sdl.com/localization-information/white-papers-articles/premium-white-papers/premium-white-papers-creating-managing-and-maintaining-a-global-website.htm

[5] Laplant M. Global content management: hewlett-packard talks the talk of worldwide business[EB/OL]. Suunyvale, USA: Hewlett-Packard, 2005 [2006-02-13]. http://gilbane.com/case_studies/HP_case_study.html

[6] AALST W, HEE K. Workflow management: models, methods, and systems[M]. Massachusetts, London: The MIT Press Cambridge, 2002: 4 - 5

[7] FISCHER L. The workflow handbook 2003[M]. FL, USA: Future Strategies Inc, 2003: 146 - 158

(责任编辑:赵海容)