文章编号:1000 - 582X(2006)07 - 0073 - 04

全球化内容管理系统框架的研究与设计*

吴国仕[1],傅湘玲[1],艾莉莎[2]
(1. 北京邮电大学 软件学院,北京 100876;2. 信息产业部 电信科学技术研究院,北京 100083)

摘 要:随着企业国际化进程的发展,中国的企业也将面临着多语言内容管理的挑战. 基于现有的企业多语言内容管理的现状,研究设计了一种能够有效集成现有的企业多语言内容管理系统框架. 首先研究了该类系统的性能特征,然后详细描述了系统的基本框架和系统的工作原理,最后通过该系统框架的实现,证明了该框架设计的可行性.

关键词: 内容管理;内容全球化;多语言内容管理;内容一致性;工作流
中图分类号:TP311.5 **文献标识码**:A

当今的世界是信息爆炸的时代,信息每天都在以惊人的速度增长. 信息的种类也在不断地扩展, IBM 研究部门调查发现,全球 85% 的信息是非结构化的,包括纸上的文件、报告、视频和音频文件、照片、传真件、信件等[1-2].

将信息有效的本地化和全球化正成为全球化经济环境下企业生存和发展的关键因素之一,而全球化的企业离不开全球化的企业信息,尤其在企业的信息内容不断增加并需要用多语言沟通时,企业必然面临一个新的挑战:如何在保持全球统一的企业品牌的前提下,开发、利用和管理多语言、多区域及多文化的企业信息. 国际知名公司成功地进入包括中国在内的全球市场的关键因素之一,是他们及早地认识到企业信息的全球化和本地化的重要性并采用了相应的先进技术加以管理[3]. 今天,跨国公司所得收入的 40% 到 60% 来自于全球市场,而这些收入的产生来自于公司有效地将它们的市场及产品信息传播到客户的能力[4]. 而将这种信息内容应用到全球市场上的过程称为企业内容全球化.

1 中国企业全球化内容管理的现状

当前,全球化的内容管理主要在一些国际公司的国内分公司以及中国本土的一些意欲国际化的企业中应用,它们实施内容全球化管理主要方式有如下 4 种:

1)按文件方式存放,各种类型的内容文档存放在同一个或不同的目录下,而对应的其他自然语言的内容文档放在另一个目录下或不同的机器上.

2)将不同语言的内容文档存放在同一个或多个数据库中,而在对应的不同语言的内容文档间没有建立任何有机的联系.

3)使用了传统的 CMS(*content management system*,内容管理系统),虽然有些 CMS 系统可以存储多语言内容信息,但这些 CMS 系统在不同的多语言内容间没有建立有效的关联机制,以至于当一种语言的内容发生了变化,另一种语言和它对应的内容不能得到及时的更新.

4)使用了混合内容存储方式,那就是上述 3 种方式的混合.

综合上述现状可以看出,现有的企业内容管理方式远远不能满足企业全球化的需要,在一定程度上制约了企业国际化的发展.

当然,全球化内容管理,并不是否认传统的企业内容管理系统的作用,而是认为企业内容全球化管理系统同企业内容管理系统的有机组合,将会加强企业的信息化建设,提高企业的管理效率,加速企业的国际化建设.

2 全球化内容管理系统的特征分析

企业信息全球化是当今全球化企业的重要标志之一,而内容管理全球化又是企业信息全球化不可缺少

* 收稿日期:2006 - 03 - 18
基金项目:信息管理与信息经济学教育部重点实验室开放基金资助课题(F0607 - 22)
作者简介:吴国仕(1957 -),男,安徽阜阳人,北京邮电大学教授,硕士,主要从事企业信息化、多语言信息管理及智能信息检索的研究.

的一个重要组成部分,提供一个能够对企业内容全球化进行有效管理的软件系统将对推动企业信息全球化具有很大的帮助作用.

按照企业内容全球化的作用及其需求,笔者认为,企业内容全球化管理系统应具有如下特征.

2.1 内容的安全性

企业内容是企业的重要内部信息资源,它包含着大量的内部信息及商业机密,因此安全性能是企业信息全球化管理系统的重要指标之一.这种安全性能应当包括身份验证和权限验证.身份验证阶段使用登录帐户来识别用户,如果身份验证成功,用户就连接到系统.然后,用户需要具备内容数据的访问权限,称为权限验证,它确定每个用户的角色并根据其角色确定用户的权限,比如读内容的权限,写内容的权限,创建新内容的权限及创建新项目的权限.

2.2 内容处理过程的自动化

企业信息全球化的过程是一个复杂的过程,单纯的手工操作是枯燥和繁琐的,而且是容易出错的,因此,信息全球化的过程应该是尽可能采用自动化过程.而由于各个企业的环境不同,管理流程也不尽相同,这就要求给用户提供用户客户化工作流程的环境,而满足上述要求最好的办法就是提供用户一套带有流程编辑功能的工作流管理系统[5].

2.3 多语言内容的相关性

所谓多语言内容的相关性是指不同语言的内容之间的关联性.这种关系是建立多语言内容一致性的关键,也是企业信息全球化管理系统区别于传统的内容管理系统的关键特点之一.

本文中的内容相关性的定义如下:

假设C是一个内容信息单元,C(lc,cc,text_id,text_content)表示C具有lc,cc,text_id,text_content 4个属性.

若C1是一个源语言内容单元,C1(lc1,cc1,text_id1,text_content1)

C2是一个目标语言内容单元,C2(lc2,cc2,text_id2,text_content2)

如果(text_id1 = text_id2) &&((lc1 ≠ lc2) || (cc1≠cc2)),我们称C1与C2具有关联性.

这里,text_id表示内容信息单元的ID,text_content代表内容信息单元的值,lc表示内容信息的语言代码,cc表示内容信息的国家代码.

2.4 内容变化的可监测性

内容变化的可监测性是指对所管理的内容信息进行全面监视,当源语言的内容发生变化,系统应该能及时发现这种变化并向系统管理人员发出提示,然后可以根据系统管理人员的指令,采用手动或自动方式启动内容全球化工作流程,进而去刷新对应的目标语言的内容.

2.5 内容的一致性

内容的一致性是指系统具有保持多语言内容一致的特点,这种特点是基于内容的相关性与内容变化的可监测性.如跨国企业的产品发布网站,当一种语言的网站发布的产品性能、价格、或销售方式进行更新或发生变化时,所对应的其他语言网站发布的内容也必须保持同步刷新.否则的话,会影响企业的形象及给企业的产品销售带来混乱.

2.6 适应环境的灵活性

不同的企业具有不同的企业文化和不同内容管理模式,因此企业内容全球化管理系统应具有较大的灵活性以适应企业的千变万化的环境需求,这种需求包括:

支持多种操作系统,如WINDOWS,UNIX,LINUX等.

支持多种文件格式,诸如DOC,PDF,TEXT,XML等.

支持多种内容存储系统,比如文件系统,数据库系统,传统的CMS系统.

支持多种内容存储形式,如内容的集中存储(同一台机器)及分散式存储(不同的机器)等.

3 全球化内容管理系统的主框架及工作原理

3.1 系统的主框架描述

系统主框架如图1所示,它包括3个主要部分:

1)企业内容全球化服务器端(称主服务器端).主服务器端主要包括应用服务器、WEB显示层、API、业务逻辑管理、项目管理、工作流管理系统、常用任务模块库及用户自定义模块库.

a.应用服务器.支持TOMCAT,JBOSS或其他WEB SERVER,主要负责处理客户端的请求及提供用户对业务流程的监视.

b.显示层.主要提供用户请求的有关界面,包括项目管理界面、语言管理界面、资源管理界面、系统管理员界面及工作流管理界面.

c.API.提供内容适配器端的接口调用及其它应用端的接口调用.

d.业务逻辑管理.是中心控制模块,这个模块根据系统配置文件及用户请求,遵循内部逻辑创建适当的翻译项目,协调服务器端各子系统及其他模块之间的关系.

e.项目管理系统.根据用户请求管理用户创建的翻译项目,一个项目需要一个项目控制文件及一个或多个翻译源文件.

f.工作流管理系统[4].主要负责管理用户创建的工作流过程的运行与调度,详细的描述见3.3全球化

内容管理系统的工作流系统框架.

g. 常用任务模块库. 提供在工作流执行过程中常用到的功能模块，如翻译预处理模块，接收翻译文件模块，翻译文件后置处理模块等. 在工作流建模过程中由流程创建人员将工作流程中的每一个任务与库中的功能模块绑定在一起，以保证工作流程的业务逻辑的具体实现.

h. 用户自定义模块库. 当通用模块库中的模块不能满足用户的特殊需求时，用户可以自定义自己的功能模块库，工作流管理系统支持工作流任务与用户定义模块之间的绑定.

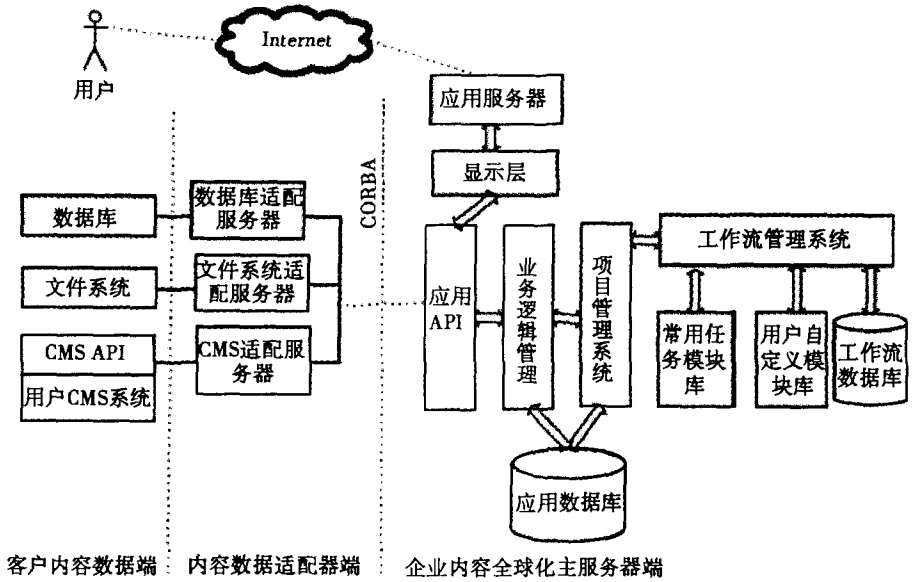

图1　企业内容全球化管理系统结构图

2)内容数据适配服务器端(称适配器端). 该部分主要负责抽取客户内容数据端的数据，并结合主服务器端进行客户内容数据的变化监测，一旦发现客户内容数据端的数据有变化，它将发送请求到主服务器端. 内容适配器端包括3种类型的适配器，它们是：

a. 数据库适配服务器. 又称数据库适配器，这种适配器以SERVICE方式运行，通常它与用户的内容数据安装在同一台机器. 它适用于客户内容数据存储在数据库的情况.

b. 文件系统适配服务器. 又称文件适配器，文件系统适配服务器亦以SERVICE方式运行，通常它与用户的内容文件系统安装在同一台机器.

c. CMS适配服务器. 又称CMS适配器，它以SERVICE方式运行，通常它与用户的CMS系统安装在同一台机器.

3)客户内容数据端. 客户内容数据端是客户所拥有的多语言数据的存放处，它的存储结构应当支持多语言内容信息的存放，它支持数据库、文件及CMS 3种数据存储格式.

3.2　系统工作原理

该系统工作原理是利用灵活的工作流管理系统与内容信息的监视系统，实现多语言信息的有效同步从而保证多语言信息的一致性. 它是一个多服务器的，多层次，多语言支持的，基于CORBA协议的系统.

适配服务器的作用主要是负责监视客户端内容信息的变化；同时负责上传需要翻译的内容信息. 一旦内容信息变化及有新内容需要翻译，它向主服务器发送请求信号，得到主服务器的批准后它将内容信息上传到主服务器. 当主服务器接到适配器的信号后，启动该类内容信息所对应的工作流程及进行相关信息的处理.

如果用户的内容信息存放在数据库中，与之对应的数据库适配器将被使用. 数据库适配器的主要功能是定期的连接用户数据库与主服务器的应用数据库，将用户数据库与主服务器应用数据库的对应的内容单元进行比对，如果发现某些内容单元发生变化，它首先增量抽取这些变化的内容，然后以依预先规范的格式(XML Schemas)[6]将这些变化的内容存储到一个XML文件中，随后它发送一个请求到主服务器，在得到主服务器同意后，它将发送一个项目控制文件和XML内容信息文件到主服务器端，由主服务器启动该文件所对应的工作流程，由工作流程将变化的内容送到翻译人员，同时，主流程将用户数据库中的变化的内容的有关信息存储到主服务器的应用数据库，并为下一次的数据变化监测做好准备. 一旦该XML文件被翻译完成，主服务器将翻译成其他语言的包含内容信息的XML文件送回到数据库适配器端，由数据库适配器解析该XML文件，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.

如果用户的内容信息是存储在文件系统中,与之对应的文件适配器将被使用.文件适配器定期地对用户的源语言文件与主服务器本地数据库中存储的该文件的信息进行比对,如果发现用户的源文件内容发生变化,它就发送一个请求给主服务器,在主服务器同意后,它将内容发生变化的文件及一个项目控制文件送到服务器端,然后启动对应的工作流程,由工作流程将变化的内容送到翻译人员,同时,它将该项目包含的文件的特征信息存储到主服务器的应用数据库,为下一次的数据变化监测做好准备.当文件被翻译完成,主服务器将新翻译的内容文件送到文件适配器,并由文件适配器去刷新原有的目标语言内容文件.

对 CMS 适配器,它定时监视用户 CMS 中的内容信息的变化.周期性地利用客户端原有的 CMS 系统的 API,抽取它所需要监控的内容文件,随后的过程类似于文件系统适配器,惟一的不同是 CMS 适配器刷新翻译后的内容文件是通过调用 CMS API 实现的.

3.3 全球化内容管理系统的工作流系统框架

工作流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(Workflow Management Coalition)标准[5],能方便地同企业现有的 JAVA,C++软件模块集成,并具有过程版本控制、资源管理、角色管理、权限管理等功能.它主要包括工作流服务器、流程设计器、流程监视器、工作流系统管理员及客户应用等模块,如图 2 所示.

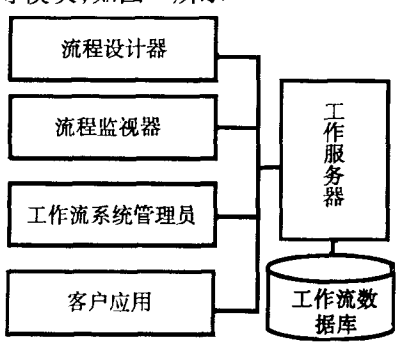

图 2 工作流系统框架

a.工作流服务器.它包括工作流引擎、资源管理、任务管理、角色管理、权限管理等模块.

b.流程设计器.提供一个友好的用户界面供客户能够方便地编辑工作流流程,对于工作流流程中的每一任务,用户还可以通过该编辑器设定其与常用任务模块库及用户自定义模块库的功能模块进行绑定,并设定该功能模块的入口参数、出口参数及完成该任务所需要的资源.一旦过程编辑完成,用户可以方便地对该工作流流程进行有效的验证[7].

c.流程监视器.监视工作流流程的实时运行状态,并提供用户参与控制运行中的工作流程的界面.那就是说,用户可以通过该界面,知道当前系统中正在运行的工作流程的状态,用户可以强制停止或启动某个正在运行的工作流程.用户还可从工作流程的一个任务跳转到另一个任务.

d.工作流系统管理员.负责管理工作流系统用户的创建,角色的创建,工作流流程的创建与管理,权限管理等.

e.客户应用.提供有关用户自定义应用的接口及有关的 API.

4 结束语

基于该系统框架的"全球化内容管理系统"目前已得到实施,该系统能对分别存放于数据库及文件系统的中、英、日文内容信息进行有效的管理.实践证明,该系统能有效地监视原语言内容信息的变化,并将变化的原语言信息送到主服务器端并由服务器端送到有关翻译部门,实现多语言内容信息的同步.同时,该系统还能减轻多语言信息 IT 管理人员的复杂的手工操作,避免了手工操作中的误操作.

随着中国企业国际化进程的发展,企业对内容管理全球化的需求越来越强烈,而研究和开发能够满足这种需求的系统管理工具对加快内容信息全球化是非常必要的和有重要意义的.笔者系统地研究了全球化内容管理系统应具有的性能特征,并在此基础上,设计了一个能够有效地集成企业现有的多语言、多种内容存放形式、多操作系统、多文件格式、易于客户化的全球化内容管理系统框架.这对企业开发自己的全球化内容管理系统及实施多语言内容信息的管理都具有一定的实用意义.

参考文献:

[1] 张 婵, 罗 佳. 企业内容管理综述[J]. 现代计算机, 2005,27(8):17-19.

[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.

[3] 方堃. 解析"企业信息全球化管理"[EB/OL]. http://www.mbaedu.cn/shtml/SXY/2005@05/16/091808.shtml, 2005-05-18.

[4] MARY LAPLANT. Global Content Management: Hewlett-packard Talks the Talk of Worldwide Business[EB/OL]. http://gilbane.com/case_studies/HP_case_study.html, 2005-01-10.

[5] FISCHER L. The Workflow Handbook 2003[M]. FL USA: Future Strategies Inc, 2003. 146-158.

[6] W3C. W3C Technical Reports and Publications[EB/OL]. http://www.w3.org/TR/, 2002-07-25.

[7] AALST W, HEE K. Workflow Management: Models Methods and Systems[M]. Massachusetts London: The MIT Press Cambridge, 2002. 4-5.

(下转第 79 页)

则依旧 P_i 用 q 代替；若 P_i 与 q 的曲率变化也很大，则插入一个点 q. 如此一来，就得到曲线上一系列点，这系列点有这样一个特点，它们包含了边缘上曲率变化大的点以及曲率大的点，同时，较均匀地 n 等分了边缘曲线，把新的点列依旧记为 $P_i(i=0,1,2,\cdots,m)$.

在许多应用领域，要求能很快速地提取目标的外形，然后进行识别，因而，根据这些点 $P_i(i=0,1,2,\cdots,m)$，采用前面快速曲线造型的方法构造出曲线：

$$\overset{\omega}{s}(t)=(x_i(t),y_i(t)) \qquad i=0,1,2,\cdots,m$$

这里，一开始的曲率只是用来判断边的弯曲程度，并不是真正要求出曲率. 所以，可以根据以下方法进行：为了计算 P_i 处曲率，用该点的前后两点间连线 $P_{i-1}P_{i+1}$（称弦）与该点到连线的距离 d_i 的长度比作为近似曲率，即 d_i/L_i，当其曲率大于某一阈值 T_g 的点，被认为是角点，即作为 q 点.

这个做法有以下好处：首先，数据量少，只需知道 $P_i(i=0,1,2,\cdots,m)$ 便可；其次，克服了多边形不能描述边缘弯曲的程度等缺点，抗干扰好，可以用于对图形的匹配分析.

3　曲线造型的边缘特征值描述

有以上方程后，可以作边缘特征值描述了，只是针对几个简单的描述符进行讨论.

3.1　边界的长度

边界的长度是一种简单的边界全局特征，它是边界所包围区域的轮廓的周长.

以往的做法要判断区域 R 的内部点是按什么方向连接的，然后确定出内部点. 然而会出现相应的歧义，即有些点在不同方向连接时，会属于内部点或边缘点，因而，按链码表示方法实际上是有 2 个边界的长度.

这里的边界长度为：

$$L=\oint_l\left|\frac{\mathrm{d}\overset{\omega}{s}(t)}{\mathrm{d}}t\right|\mathrm{d}t,$$

其中 l 表示边缘一周进行积分.

于是

$$\left|\frac{\mathrm{d}\overset{\omega}{s}(t)}{\mathrm{d}t}\right|=\left[\left(\frac{x_i-x_{i+1}}{t_{i+1}-t_i}+bx\frac{(t-t_i)(t-t_{i+1})}{(t_{i+1}-t_i)(t_i-t_{i+1})}+\left(\frac{t-t_i}{t_{i+1}-t_i}a_x+b_x\right)\frac{2t-t_i-t_{i+1}}{(t_i-t_{i+1})(t_{i+1}-t_i)}\right)^2+\left(\frac{y_i-y_{i+1}}{t_{i+1}-t_i}+b_y\frac{(t-t_i)(t-t_{i+1})}{(t_{i+1}-t_i)(t_i-t_{i+1})}+\left(\frac{t-t_i}{t_{i+1}-t_i}a_y+b_y\right)\frac{2t-t_i-t_{i+1}}{(t_i-t_{i+1})(t_{i+1}-t_i)}\right)^2\right]^{1/2}$$

当然，如此做较为复杂，可以用下式进行计算：

$$\left|\frac{\mathrm{d}\overset{\omega}{s}(t)}{\mathrm{d}t}\right|=\left|\frac{x_i-x_{i+1}}{t_{i+1}-t_i}+bx\frac{(t-t_i)(t-t_{i+1})}{(t_{i+1}-t_i)(t_i-t_{i+1})}+\left(\frac{t-t_i}{t_{i+1}-t_i}a_x+b_x\right)\frac{2t-t_i-t_{i+1}}{(t_i-t_{i+1})(t_{i+1}-t_i)}\right|+\left|\frac{y_i-y_{i+1}}{t_{i+1}-t_i}+b_y\frac{(t-t_i)(t-t_{i+1})}{(t_{i+1}-t_i)(t_i-t_{i+1})}+\left(\frac{t-t_i}{t_{i+1}-t_i}a_y+b_y\right)\frac{2t-t_i-t_{i+1}}{(t_i-t_{i+1})(t_{i+1}-t_i)}\right|.$$

3.2　曲率

曲率是斜率的改变率，它描述了边界上各点沿边界方向变化的情况.

根据曲率的定义：

$$k(t)=\left|\frac{\mathrm{d}\overset{\rho}{s}(t)}{\mathrm{d}t}\times\frac{\mathrm{d}^2\overset{\rho}{s}(t)}{\mathrm{d}t^2}\right|\Big/\left|\frac{\mathrm{d}\overset{\rho}{s}(t)}{\mathrm{d}t}\right|^3,$$

同样，可以获得曲线的曲率.

当然，由于有了边缘点的近似曲线方程，故像傅里叶描述符、区域面积、形状描述符等内容均可计算出来，因而有很广泛的应用.

4　实验与分析

按照此思路进行编程实现，程序用 Visual C++ 6.0 编写的，在 Windows2000 操作系统，主机频率为 1.5 GHz下完成的，图 2 为画出的曲线情况，由此可见，效果较好，边缘曲线描述很准确，同时运行速度极快. 在图形匹配、图像视觉处理等领域有广泛地应用价值. 当然，同许多 B 样条曲线一样，这种曲线造型无法描述直线等常见二次曲线. 因此，以后研究中会考虑如何形成快速 NURBS 曲线来进行分析，克服这个困难.

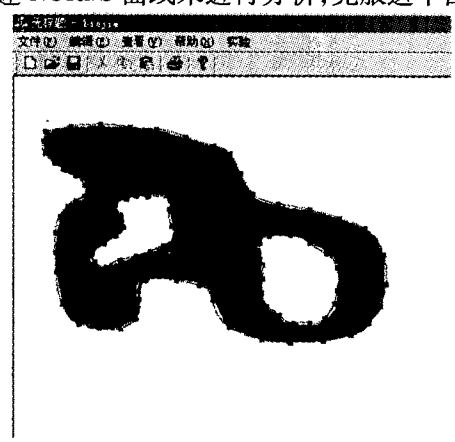

图 2　用自由曲线描述边缘线

参考文献：

[1] 林意，袁琦睦，何援军. 一种过型值点的快速曲线造型算法[J]. 中国工程图学学报，2005，(4)：72－76.

[2] 章毓晋. 图象处理和分析[M]. 北京：清华大学出版社，1999.

[3] 杨平，林意. 一种三角剖分算法实现图形的匹配[J]. 计算机应用与软件，20(2)：70－71.

[4] 袁琦睦，林意. 一种单值区域的图形匹配算法[J]. 计算机应用与软件，2005，22(8)：101－102.

[5] 林意. 过控制顶点的 B 样条曲线[J]. 江南大学学报，2003，(6)：553－556.

[6] 林意，熊汉伟，骆少明，等，过控制顶点的B样条曲线[J]. 江南大学学报，2003，2(6)：553－556.