

基于 Web 的新闻自动采编系统^{*}

张 巍, 陈三忠

(深圳大学网络中心, 广东 深圳 518060)

摘 要: 为了适应自动采编网络新闻的应用需求, 提出了一种基于 Web 的新闻自动采编系统。首先给出了系统的三层体系结构, 在此基础上阐述了各功能子系统的作用和特点, 最后给出了深圳大学新闻抓取服务器的具体实现。

关键词: 新闻; Web; 自动采集

中图分类号: TP393.09 **文献标识码:** A **文章编号:** 0529-6579 (2002) S1-0057-03

随着 Internet 的发展, 网络信息量急剧膨胀, 怎样及时地采集网上新闻成为许多新闻网站和新闻监管部门的迫切要求。而传统的手工采集加工或者完全手工编辑新闻, 对于现在日益庞大的新闻量和用户最快速地获取信息的要求是不相适宜的, 因此本文提出了一种基于 B/S 结构的新闻自动采编系统。该系统不仅能满足以上需求, 而且其功能可扩展, 以提供更多的增值服务。

1 系统体系结构

新闻自动采编系统基于 B/S 结构, 客户端完全零安装, 只要有浏览器就可以了。普通用户通过浏览器访问该系统发布的新闻, 系统用户也只需通过浏览器就可以实现从远程对系统的使用和管理, 所有功能都在后台服务器中完成, 对用户完全透明。服务器端系统的体系结构如图 1 所示。

Web 服务器					
系统管理	信息采集	加工编辑	模板设计	新闻发布	……
文件系统			数据库		

图 1 新闻自动采编系统结构图

该系统共分为 3 层: Web 服务器层、功能子系统层和存储层, 其中功能子系统层分为系统管理子系统、信息采集子系统、加工编辑子系统、模板设计子系统和新闻发布子系统, 存储层综合利用了数据库和文件系统, 采集的新闻先入库存储, 对于经常访问的新闻生成静态页面, 以提高

访问的速度, 而其它的用户数据、统计数据等等也存入数据库中, 便于随时的查询和更新。另外, 在功能层中可以继续扩充其它功能子系统, 例如信息检索子系统等, 因为信息的自动采集和入库已经实现, 所以可以在此基础上增添新的功能, 建立索引实现关键字和全文检索。

2 各子系统的功能和特点

系统管理子系统包括用户管理、统计监测和系统维护升级。整个新闻系统依据不同的权限划分成不同角色, 例如: 系统管理员角色、信息采集角色、加工编辑角色等等, 然后将角色赋给不同的用户, 从而满足不同用户的权限需求, 其中系统管理员角色具有所有权限, 并负责用户的创建删除和角色的分配。系统管理子系统还对整个系统的各种采编数据进行统计报告, 供用户参考。

信息采集子系统分为配置和后台采集两部分。用户首先按照需要设置不同的新闻类别, 例如国际新闻、国内新闻等, 在不同类别中分别确定不同新闻来源, 即从计划采集新闻的网站上确定出包含相关新闻列表的网页的网址, 然后分析该网页的源文件, 确定出每条新闻的绝对网址, 接着通过这些绝对网址逐一获取新闻内容页面, 分析其源文件, 确定新闻标题、正文、图片网址等, 最后将标题、正文、图片等全部抓取到本地数据库中。一般新闻网站的新闻列表和新闻内容页面的格式相对固定, 对于页面中试图抓取的信息, 分析出其前后可惟一确定它们的 html 标记作为

^{*} 收稿日期: 2002-03-21

作者简介: 张 巍 (1976-), 女, 助理工程师, E-mail: zhangwei@szu.edu.cn

特殊标识, 从而通过程序自动地采集该网站上的相关新闻信息。这是自动采集的基本原理。后台采集的程序可以依照不同要求定时执行, 及时的获取各大新闻网站的新闻。但是, 一旦信息源的格式发生变化, 系统就要及时提醒用户更改相应配置。新闻采集的基本流程如图 2 所示。

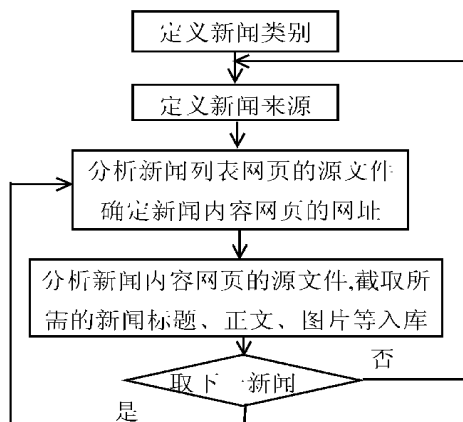


图 2 新闻采集流程

信息采集子系统只是简单的抓取其它网站上的新闻, 并不一定符合用户的需求, 因此通过加工编辑子系统, 既可以对采集的新闻内容进行加工修改, 甚至删除, 也可以增加某些功能, 例根据新闻的点击率统计出每日热点新闻、每周热点新闻等, 还可以手工编辑自己的新闻存入数据库中。

信息采集和加工编辑子系统主要针对新闻的内容进行处理, 而浏览新闻的普通用户所看到的网页的界面由模板设计子系统确定。将新闻内容和网页的表现形式分开, 从而能按照不同的需要设计不同的版面。无需改变采编程序的代码, 只用修改相应的模板设计, 新闻网页的改版更加方便迅速。

一般新闻发布系统的做法有两种: 一是把新闻内容直接生成 html 文件, 按日期和新闻 ID 号来命名文件, 一些大的新闻网站常这样处理; 另一种就是把新闻信息保存到数据库里, 当用户阅读新闻时, 从数据库中调出数据, 动态的生成页面。前一种方法显示速度快, 服务器负荷小, 后一种方法维护方便, 处理简单¹。针对本系统特点, 当系统空间允许, 普通用户的新闻访问量很大时, 可以结合这两种方法。因为新闻内容和网页模板是分开的, 对于已加工好可以直接发布的新闻, 不妨临时生成新闻内容的静态网页, 在数据库中存放相应的链接地址, 方便用户访问。当网页模板更新时, 则必须采用动态网页方法, 直接从数据库中生成

用户想要访问的网页, 同时更新所有静态网页。

3 具体实现

根据新闻自动采编系统的体系结构, 深圳大学校园网中新闻抓取服务器初步实现了简单的新闻采集和发布功能。校园网的内部用户通过新闻抓取服务器就可以访问 ChinaNet 上的主要新闻, 无需再直接访问各大新闻网站, 也可以减少到 ChinaNet 的网络流量。该新闻抓取服务器建于 Redhat Linux7.1 操作系统上, 采用了 MySQL3.23.44、Apache1.3.22、PHP4.0.6 和 FastTemplate 模板等软件。

采用 PHP、MySQL 和 Apache 构筑网站是目前的流行方法之一。MySQL2 是一个免费的数据库服务器软件, 相对于其它大型数据库系统而言比较简单, 容易控制。Apache3 是目前使用最广泛的 Web 服务器程序, 它的源代码是免费的, 可运行于 Unix、Windows 平台。PHP4 是一种多平台无缝运行的服务器端嵌入式脚本语言, 它融合了 C、Java 和 Perl 等编程语言, 也是一种免费的自由软件, 可以在大多数 Unix 平台、GNU/Linux 和微软 Windows 平台上运行, 移植速度快, 且具有丰富的函数库。将 PHP4 作为 Apache 的静态模块安装, 同时还要带上对 MySQL 的支持, 这样就能利用 PHP 函数访问 MySQL 数据库了。

FastTemplate 模板用于模板设计, 它能有效的将 html 和 PHP 代码分开, 不仅在开发阶段可以分别提高界面设计人员和应用程序编写人员的工作效率, 更会给项目的测试和维护带来巨大的便利。通过网页设计工具设计新闻网页的版面, 生成相应的扩展名为 .tpl 模板文件, 然后每次赋给对应的内容, 通过函数 FastPrint() 生成网页供客户端浏览。我们正是通过 FastTemplate 来实现对新闻网站外貌的迅速改变。

后台采集功能用命令行方式的 PHP 实现, 利用 Linux 操作系统的 cron 命令来实现定时采集。下面我们通过一个例子来说明怎样分析出网页源文件中的特殊标识, 从而采集到想要的新闻信息。例如某新闻网站的国内新闻列表网页网址为 <http://www.news.net/china/>, 从其 html 源文件中截取一段新闻列表代码如下:

```
<LI><A href= /content/2002-03/10/content-289427.htm target= blank> 国内新闻标题 1 </A>
</FONT size= 12> (22; 16 10 日) </FONT>
```

<LD> 国内新闻标题 2< /A > (22: 20 10 日)< /FONT>

从中可以得出国内新闻 1 和国内新闻 2 的内容网页的绝对网址为:

http: //www. news. net /content /2002 - 03/10 / content -289427. htm

http: //www. news. net /content /2002 - 03/10 / content -289435. htm

所以, 可以将相对网址前后的可唯一区分它的代码<LD> 作为特殊标识。因为网页的格式是相对固定的, 采集程序下载了网页的源文件后, 查找特殊标识取得网页相对网址, 再同网站网址组成绝对网址, 供下一步下载新闻内容网页使用。这些特殊标识必须可唯一区分出新闻内容相对地址。新闻内容网页的分析原理与此类似, 从其源文件采集新闻标题、正文、图片等入库。采集的新闻存入 MySQL 数据库中, 相关的表结构如表 1 和表 2 所示。

表 1 新闻正文表

字段名	数据类型	字段含义
id	int	新闻标号
title	char	新闻标题
class	int	新闻类别
source	char	新闻来源
content	text	新闻正文内容
img	int	是否有图像的标志
time	date	采集时间
hits	int	新闻点击次数

新闻发布时先自动生成新闻标题列表的 javascript 文件, 每当用户访问时系统利用 FastTemplate 模板动态生成新闻列表页面, 而当用户

点击新闻标题链接时, 再动态地从数据库中取出新闻内容, 包括正文和图片, 最后利用 FastTemplate 模板动态生成新闻内容页面。同时, 系统还能根据新闻点击次数, 统计出每日、每周的热点新闻。

表 2 新闻图像表

字段名	数据类型	字段含义
id	int	图像标号
news-id	int	新闻标号
img_name	char	图像标题
img_data	blob	二进制图像
img_type	char	图像类型

4 结束语

在网上新闻每日剧增的今天, 构建出新闻自动采编系统可以减轻新闻工作者的劳动负担, 提高工作效率。文中提出的新闻自动采编系统体系结构, 具有良好的可扩展性, 能够在基本采集功能之上扩充其它相关功能, 满足用户的不同需求。基于这种体系结构实现的深圳大学新闻抓取服务器, 目前只具备基本的自动采集和动态发布功能, 还有待于进一步的完善和提高。

参考文献:

[1] 冯天飞,施法中,王瑛.基于 Web 的新闻自动发布系统的设计和实现[J]. 计算机应用,2001, 21(3): 66-67.

[2] http: //www. mysql. com

[3] http: //www. apache. org

[4] http: //www. php. net

[5] Sascha Schumann Templates— why and how to use them in PHP3 http: //www. phpbuilder. com/columns/sascha - 19990316. php3

Automatic News Gathering and Editing System Based on Web

ZHANG Wei, CHEN San zhong

(Network Center, Shenzhen University, Guangdong, Shenzhen 518060, China)

Abstract: For the need of gathering web news automatically, this paper provides an automatic news gathering and editing system based on web. First a three level artitecture of the system is proposed. Then the function and feature of every functional subsystem is discussed. Finally the implementation of news grabber server in Shenzhen University is put forward.

Key words: news; Web; automatic gathering