

基于 RSS 技术的校园统一资讯平台设计

汪文彬

(浙江科技学院 校园网管理中心,杭州 310023)

摘要: 目前校园内通知公告等信息都通过网站发布,但各个站点之间相互独立,用户要及时全面获取不同类型的信息需要频繁在各个网站之间来回穿梭。基于 RSS 技术构建校园统一资讯平台,实现了校内各站点上信息资源的自动提取与采集,以 RSS 的格式发布,同时对采集到的信息进行分类、在线聚合,可方便校内用户个性化订阅,提高信息发布效率。

关键词: RSS;资讯;聚合;采集;订阅

中图分类号: TP393.09

文献标识码: A

文章编号: 1671-8798(2009)02-0104-06

Design of campus unified information platform based on RSS

WANG Wen-bin

(Campus Network Management Center, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: The informations are issued by websites in the campus in current. Since the websites are independent to each other, consumers must go through them frequently if they want to get different informations in time. The campus unified information platform based on RSS can collect the information resources of each site in campus automatically and issue the results in RSS. Meanwhile, it classifies the collected informations and aggregates online. In this way, it is convenient for internal consumers to subscribe information individually and enhance the efficiency of information issuance.

Key words: RSS(Really simple syndication); information; aggregation; collection; subscription

随着信息技术的快速发展,信息资源数字化、网络化趋势日益明显。传统的海报、校园广播、纸质通知等已不再是校园信息传播的主要渠道,人们更多地开始借助网站来发布和获取信息。面对日益剧增的信息资源,用户一般只能借助网络搜索引擎获取自己关心的信息。但对于新闻、通知、公告这类频繁

更新的信息,通过搜索引擎方式获取的信息无法保证是最新的。因此用户要想获取最新的信息,必须遍历各个自己关注的网站获取。在一些栏目甚多的网站中,用户往往需要进行三次甚至更多的点击才能看到最终的内容页面,而且很多信息都是因为关注不够及时而过期的。

收稿日期: 2009-03-10

作者简介: 汪文彬(1980—),男,浙江淳安人,助理工程师,硕士研究生,主要从事校园信息化建设、网络安全与防范、数据中心管理维护工作。

如何协助校内用户全面、及时、高效地获取所需信息成为一种迫切需要。本文提出一种以RSS同步信息传递模式,结合XML标准,整合校内各类信息资源,构建校园统一资讯平台,能很好地解决目前在校内信息发布中存在的上述问题。

1 RSS技术

1.1 RSS简介

RSS作为信息交流方式的一种标准,起源于Netscape(网景公司)的“推”技术。其本质是实现信息聚合的技术,是站点与站点、站点与用户之间共享内容的一种简易信息发布和传递的方式。RSS的具体含义与其版本有关,最初的版本为RSS 0.90,其全称为RDF site summary(RDF站点摘要),但随着RSS 0.91版本的出现,RSS被重新命名为Rich site summary(丰富站点摘要)。在随后出台的RSS 0.92,RSS 0.93和RSS 0.94等版本,为了强调其所做的简化工作,将RSS的全称定为Really simple syndication。RSS目前最新版本为RSS 2.0。RSS文件包含了信息源网站的全部或归纳后的内容,同时也可包含传送内容的附加信息,例如:文章发表日期、作者等。RSS技术的应用既方便了信息提供方也方便了读者,内容提供方可以自动将欲发布的内容发布到订阅者的聚合器内,同时,订阅者可以将自己喜爱的网站聚合在一个RSS聚合器内,并在第一时间获得网站更新的内容。该技术目前已广泛应用于新闻、博客、Wiki、实时资讯、气象预报等^[1]。

1.2 RSS基本原理

RSS技术是基于XML标准建立的内容包装和投递的协议,它规范了网站发布更新消息时的格式,要求以XML格式记录信息的题目、作者、发布时间、摘要内容、相关的URL地址等。网站更新内容时,只需要按照RSS标准生成同样形式的文件,RSS订阅工具可以检测网站发布文件并自动地将更新的文件下载到本地。用户通过RSS reader可以浏览到相应频道给出的信息列表,通过刷新可以查看最新的频道更新内容。同时通过相应的URL地址可以链接到原始网站查看详细内容。RSS搭建了一个信息迅速传播的平台,使得用户可以不用逐个登陆网站而实时获取最新消息。RSS标准规定的XML格式文件使得文件中包含的信息能直接被其他站点调用,同时也能在其他的终端和服务中使用^[2]。

1.3 RSS体系结构

RSS的体系结构主要由内容提供者(Content provider)、RSS聚合器(RSS aggregator)和浏览器(viewer)三部分组成,如图1所示。

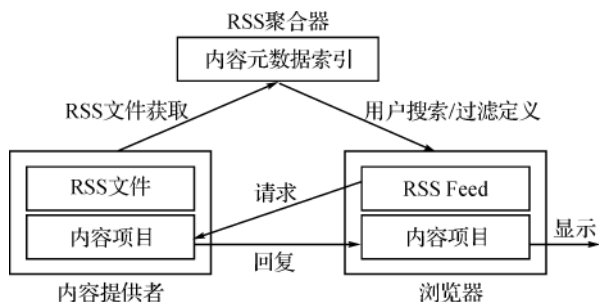


图1 RSS体系结构

Fig.1 RSS architecture

1.3.1 内容提供者 提供的内容一方面包括完整内容的页面,另一方面还要提供对该内容进行描述的RSS文件。

1.3.2 RSS聚合器 定时到众多的消息源读取最新的RSS文件,汇总并进行索引,并按索引提供读者已定制的特定主题的消息。RSS聚合器主要有在线(Centralized)和桌面(Personal)2种类型。

1.3.3 浏览器 以用户订阅为基础,标题浏览器得到用户的请求后,连接到RSS聚合器,获取文档链接源,并显示给读者。读者在浏览消息时,可以通过点击消息标题的链接,直接进入内容提供者的网站阅读详细内容^[3]。

1.4 RSS文档结构

所有的RSS文档必须遵循W3C网站上公布的XML 1.0规范。在一个RSS文档中,首先要对XML进行声明,定义文档中使用的XML版本和字符编码;根元素是<rss>,带有一个必备属性version,用以指明该文档遵循的rss规范。<rss>元素只有一个子元素<channel>,用于描述RSS feed。<channel>元素有3个必需的子元素:<title>频道的标题、<link>频道的超链接和<description>频道描述。一个<channel>元素可拥有一个或多个<item>元素,每个<item>元素可定义RSS feed中的一篇文章。<item>元素有3个必需子元素:<title>项目的标题、<link>项目的超链接和<description>项目描述。<channel>和<item>都可以分别包含若干个子元素,子元素必须成对使用。文档最后为2行关闭<channel>和<rss>元素^[4]。RSS文档结构如下:

```
<rss version="2.0">
<channel>
  <title>频道名称</title>
  <link>频道 URL </link>
  <description>频道描述</description>
  <item>
    <title>项目标题</title>
    <link>项目 URL</link>
    <description>项目描述</description>
    <author>项目来源</author>
  </item>
  .....
</channel>
</rss>
```

2 系统框架与功能分析

2.1 系统框架

目前,部分网站已经涵盖了 RSS 信息订阅的功能,但大多数站点仍然是以 HTML 静态页面或是 ASP、JSP 等动态页面的形式发布信息。系统框架包含了信息提取与采集、RSS 发布等模块,如图 2 所示。

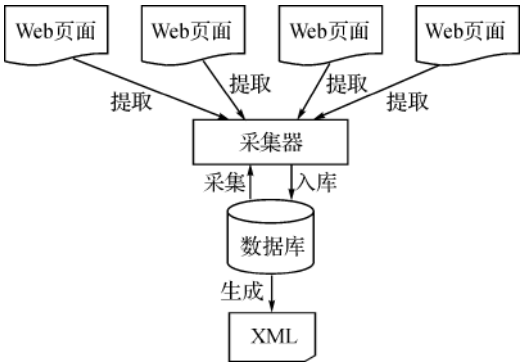


图 2 系统框架

Fig.2 Framework of system

2.2 信息采集

信息采集是本系统功能实现的基础。采集过程分三步实现：

第一步,利用 HTTP 协议,向被采集页面发送请求,得到被采集页面的 HTML 代码。通过 HTML 分析器对网页代码进行分析,准确定位信息标题、内容显示页链接、信息内容、发布时间、信息来源等,生成采集配置文件供采集器对信息进行采集,如图 3 所示。

第二步,采集器通过采集配置文件快速完成信

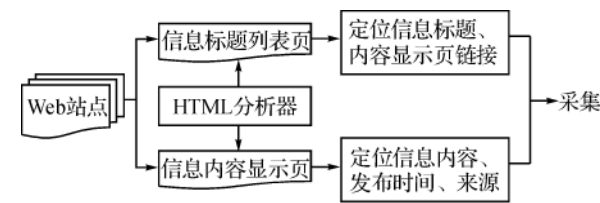


图 3 信息提取

Fig.3 Information collection

息采集。

第三步,将采集到的信息写入数据库进行存储。

2.3 以 RSS 形式发布

对于采集到的信息,如果仅仅只是以获取作为目的的话,按照有关的 HTML 格式就可以发布了。但这只是起到了聚合的作用,没有发挥订阅模式的优势。针对这种情况,在信息发布这个环节采用 RSS 标准来实现对信息的组织。由于 RSS 标准的开放性,用户可以使用 RSS 聚合器很方便地读取 RSS 文档,同时整个 RSS 文档的各个标签都定义得很明确的语义,因此在用户获取文档的过程中便能准确地获得整个文档的描述以及相关信息。

3 系统实现的关键技术

3.1 数据库设计

系统采集到的各类信息通过 SQL Server2000 数据库保存,数据库主要包括 2 个表(表 1,表 2),Channel 表用于存储待采集页面的相关信息以及采集的相关配置参数,Content 表用于存储采集器采集的内容。

表 1 Channel 表

Table 1 Channel

字段名	类型	大小	备注
ChannelID	smallInt	2	频道 ID
ClassID	smallInt	2	所属栏目
ChannelTitle	varchar	30	频道名称
ChannelLink	varchar	50	频道地址
ChannelDemo	varchar	200	频道描述
ListStartStr	varchar	200	信息标题列表开始标记
ListEndStr	varchar	200	信息标题列表结束标记
HrefStartStr	varchar	200	信息内容链接开始标记
HrefEndStr	varchar	200	信息内容链接结束标记
TitleStartStr	varchar	200	标题开始标记
TitleEndStr	varchar	200	标题结束标记
ContentStartStr	varchar	200	正文开始标记
ContentEndStr	varchar	200	正文结束标记

表 2 Content 表
Table 2 Content

字段名	类型	大小	备 注
ContentID	int	4	信息 ID
ChannelID	smallInt	2	频道 ID
Title	varchar	50	信息标题
Content	ntext	16	信息内容
Demo	varchar	200	信息描述
Link	varchar	50	内容链接
Author	varchar	20	作者
Hits	int	4	点击次数
CollectTime	smalldatetime	4	采集时间
Editor	varchar	20	编辑
OnTop	bit	1	是否置顶
Keyword	varchar	50	关键词

3.2 远程获取网页 HTML 代码

MSXML 中提供了 Microsoft.XMLHTTP 对象,能够完成从数据包到 Request 对象的转换以及发送任务。XMLHTTP 是个传送 XML 格式数据的超文本传输协议。利用 XMLHTTP 进行数据传输,上传的指令和下达的结果可以是 XML 格式数据,也可以是字符串、流,或者是一个无符号整数数组,还可以是 URL 的参数。客户端调用 XMLHTTP 的过程大致分为 5 个步骤:

- 1) 创建 XMLHTTP 对象;
- 2) 打开与服务端的连接,同时定义指令发送方式,服务网页(URL)和请求权限等;
- 3) 发送指令;
- 4) 等待并接收服务端返回的处理结果;
- 5) 释放 XMLHTTP 对象^[5]。
- 实现代码如下:

```
Function GetHttpPage(ByVal URL, ByVal Cset)
    Dim BlockStartTime
    On Error Resume Next
    Dim Http
    If isNull(URL)=True or Len(URL)<
18 or URL="MYMFalseMYM" then
        GetHttpPage="MYMFalseMYM"
        Exit Function
    End if
    BlockStartTime=Timer()
    Set Http=server.createObject("MSXML2.
XMLHTTP")
```

```
Http.open "GET",URL,False
Http.Send()
Dim temp,BlockTimeout
BlockTimeout=64
While(http.ReadyState <> 4)
    temp=Timer() - BlockStartTime
    Response.Write(Timer())
    If (temp>BlockTimeout) then
        http.abort
        Set Http= Nothing
        GetHttpPage="MYMFalseMYM"
        Exit function
    Response.End
End if
http.waitForResponse 10000
Wend
If Http.Readystate<>4 then
    Set Http= Nothing
    GetHttpPage="MYMFalseMYM"
    Exit function
End if
GetHTTPPage= bytestoBSTR (Http.
responseBody,Cset)
Set Http= Nothing
If Err.number<>0 then
    If IsNull(URL)=True or Len(URL)<
18 or URL="MYMFalseMYM" then
        GetHttpPage="MYMFalseMYM"
        Exit Function
    End if
    Set Http= Nothing
    Err.Clear
End if
End Function
```

3.3 生成 RSS feed

RSS feed 就是 RSS 的 Web 内容源,有人叫它“种子”,其实质是基于 XML 的一种文档、代码^[6]。生成 RSS feed 的基本流程是:根据 RSS 文档结构从数据库中循环读取采集到的数据,然后将相应的字段内容依次匹配到 RSS 文档结构标记中,格式化为符合 RSS 定义的 XML 标记并生成 XML 文档。RSS feed 文件在浏览器中显示效果如图 4 所示。

```

<? xml version="1.0" encoding="utf-8"? >
<? xml-stylesheet type="text/xsl" title="XSL Formatting"
href="ZUSTRSS.xsl"? >
<rss version="2.0">
<channel>
<title>浙江科技学院通知公告 RSS</title>
<image>
<title>浙江科技学院通知公告 RSS Feed</title>
<link>http://news.zust.edu.cn/? infotype=00001</link>
<url> http://news.zust.edu.cn/images/zustrssfeed.jpg</url>
</image>
<link>http://news.zust.edu.cn/? infoType=00001</link>
<copyright>浙江科技学院</copyright>
<description>RSS </description>
<item>
<title>关于元旦放假期间教工班车安排的通知</title>
<link>http://news.zust.edu.cn/Notice.asp? ID=455</link>
<author>产业后勤管理处</author>
<pubDate>Wed, 24 Dec 2008 11:28:22 GMT</pubdate>
<description>根据学校关于元旦放假安排的通知.....
</description>
</item>
</channel>
</rss>

```

图 4 RSS Feed 文件(摘录)

Fig.4 RSS Feed(excerpt)

3.4 RSS 在线聚合

RSS 订阅的方式可以通过客户端软件,如遨游 RSS、Foxmail RSS 阅读器等,此方式需要使用者本地下载安装阅读器。搭建在线聚合器随时随地地实现频道订阅、信息推送、信息聚合、共享等功能,用户只需登录在线聚合器,就能直接浏览各订阅网站发布的最新信息,不必逐一登录不同的站点,从而节省时间和费用。聚合器需要通过 XMLReader 类读取 xml 文件,实现代码如下^[6]:

```

Function XMLReader (RssSource)
Dim objXMLHTTP, strReturn
Set objXMLHTTP=server.CreateObject("
Microsoft.XMLHTTP")
objXMLHTTP.Open "Get", RssSource,
False
objXMLHTTP.setRequestHeader "Content-Type", "text/xml"
objXMLHTTP.setRequestHeader "charset", "
UTF-8"
objXMLHTTP.Send

```

浙江科技学院 • RSS Feed

浙江科技学院 通知公告RSS

RSS是站点用来和其他站点之间共享内容的简易方式(也叫聚合内容)。RSS使用XML作为彼此共享内容的标准方式。它代表了Really Simple Syndication (或RDF Site Summary, RDF站点摘要)。

关于元旦放假期间教工班车安排的通知

根据学校关于元旦放假安排的通知精神,2009年1月1日至3日放假期间的教工班车按周日的班车时刻表运行,1月4日(星期日)的教工班车按周一至周五的班车时刻表运行.....

[查看详细内容](#)

北京发布时间:2008年12月24日 19:28:22 (距离现在92天1小时1分钟)

关于举办优秀企业家进校园(第27期)报告会的通知

周建中,中冠建筑装饰工程有限公司总裁,紫蝶名家总裁,中国建筑装饰协会常务理事,中国美院客座教授,紫蝶间品国际设计机构总设计师,浙江省建筑装饰行业协会副会长.....

[查看详细内容](#)

北京发布时间:2008年12月18日 19:25:52 (距离现在98天1小时4分钟)

图书馆关于再次开展免费代查代检馆外文献服务的通知

为了更好地满足我院师生对本校图书馆收藏以外的文献的需求,图书馆再次推出免费的文献代查代检服务。有需要的读者可以在图书馆主页“文献传递”栏目里下载相关表格,发送到图书馆咨询部邮箱.....

[查看详细内容](#)

北京发布时间:2008年12月18日 19:21:55 (距离现在98天1小时8分钟)

```

testXML=objXMLHTTP.responseText
Set objXMLHTTP=Nothing
set xmldoc=Server.CreateObject("MSXML2.
DOMDocument")
xmldoc.loadXML(testXML)
Set Root=xmldoc.documentElement.selectSingleNode("/rss/channel/item")
For Each nodeX in root.selectnodes("/rss/channel/item")
Response.Write("<a href=" & NodeX.
childNodes.Item(1).Text & "' target='_blank' alt
=" & NodeX.childNodes.Item(2).Text & ">"
& NodeX.childNodes.Item(0).Text & "</a><br
>")
Next
Set xmldoc=Nothing
End Function

```

统一资讯平台页面根据 XMLReader 类依次读取 xml 文件并在浏览器中显示,如图 5 所示。



图5 RSS资讯平台

Fig.5 RSS information platform

4 结 语

本系统主要针对目前在校园内信息发布中存在的诸多不便,提出了校园统一资讯平台,实现了对各类信息的采集和聚合。应用RSS技术实现快速、有效地发布和检索信息是信息发布的一个新趋势,它对信息的实时更新,不仅能节省成本,还能提高信息发布的效率。当然,该系统目前只能说是一个雏形,还有很多地方值得进一步深入研究与开发。

参考文献:

[1] 胡智文.RSS与语义网研究[J].计算机工程与设计, 2008,29(17):4618-4620.

[2] 谢倩堃.RSS新闻的更新特征分析及RSS Reader的订阅模型[D].北京:北京交通大学电子信息工程学院, 2008.

[3] 周艳,陈永建.基于RSS技术的信息发布系统设计与实现[J].北京联合大学学报:自然科学版,2008,22(4):40-43.

[4] 佚名.RSS2.0规范[EB/OL].(2004-04-07)[2009-01-20]. <http://www.donews.net/softbunny/articles/11030.aspx>.

[5] 张俊灵,李也白,李天立,等.基于XMLHTTP组件实现电子政务平台下的站内消息系统[J].计算机与信息技术,2007,16-18.

[6] 许新华.ASP+SQL Server环境下RSS Feed的程序实现[J].湖北职业技术学院学报,2005,8(4):65-67.