# 基于 Web Services 和插件架构的新闻中心平台的设计

李 剑[1],陈海建[2,3]

(1.江苏广播电视大学,江苏 南京 210036;
2.上海电视大学 信息与工程系,上海 200433;
3.上海财经大学 信息管理与工程学院,上海 200433)

**摘 要**:设计并实现了一个网页插件形式的半自动化的网页新闻信息搜集中间件,该中间件能将关注网站的信息搜集并集中。此外,基于该中间件,还设计了一个新闻信息的汇集平台,并以江苏城市职业学院网站作为应用实例展示了校园新闻中心。

**关键词**:Web Services;架构;新闻中心

**中图分类号**:TP319 **文献标识码**:A **文章编号**:1674-7720(2012)23-0014-04

## The design of media centre based on Web Services and plug-in framework

Li Jian[1], Chen Haijian[2,3]

(1.Jiangsu Radio and TV University, Nanjing 210036, China;
2.Department of Information Technology, Shanghai TV University, Shanghai 200433, China;
3.Department of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China)

**Abstract**: The paper illustrates the design of a Web plug-in which is used for collecting and classifying information from websites. Taking the website of JSCVC as an example, the paper also illustrates the application of the plug-in.

**Key words**: Web Services; Plug-in; media center

随着网络技术的迅猛发展,以网页形式发布的信息、数据爆炸性地增长,人们期望能够及时有效地关注、获取、管理和分析对比特定行业的热点信息和数据。尤其在金融和传媒领域,信息和数据变化特别快,能及时收集来自不同网站上发布的信息和数据变得尤其重要。但是,绝大多数网页基于 HTML 格式,其松散的特性使得现有的页面提取算法不能较好地与特定领域的业务相结合,分析出信息的语义。例如,对单个页面而言,如果完全进行自动化分析,较难在分析结果中对与新闻这一特定领域相关的新闻标题、链接、发表时间和正文等内容元素加以区分;对于新闻标题列表和新闻正文分别位于不同网页这一典型的多页面结构,需要联合分析多个网页的内容,如果没有一套完善的提取和分析这些信息的软件平台,也难以得到完整信息。

本文针对 Internet 上新闻网站的信息,设计并实现了一个网页插件形式的半自动化的网页信息搜集中间件,对于不同版式的新闻网站,只需编写少量相关代码实现中间件,即可完成对新闻标题、链接、发表时间和正文等信息的提取。新闻生成器插件加入到新闻中心平台中,由 Web Services 提供同构的新闻获取接口,前台通过调用 Web Services 即可构建汇集各种新闻和通知公告的新闻中心。此外,本平台还提供将新闻转换为 RSS(Really Simple Syndication)种子的功能,以方便用户使用 RSS 阅读器订阅新闻。

基于此种新闻中心平台架构,完全不需要改变现有的新闻网站布局,即可自动汇聚多方面的新闻,创建新闻门户网站。此外,本平台还具有高扩展性,当有新的网站需要关注时,只需编写扩展插件,即可实现将关注的网站加入到视野中。

## 1 网页信息提取技术

对 Web 信息提取的研究早在 20 世纪 80 年代就已开始,根据参考文献[1],现有的对 Web 信息提取分析的方法可以分为多种类型:从自动化程度上可划分为手工、半自动和全自动提取分析方法;从原理上可划分为

基于自然语言理解、基于本体、基于 HTML(HyperText Markup Language)和基于隐马尔可夫模型等提取分析方法。本文设计并实现的新闻中心平台,在对单个页面进行分析时所采用的是基于 HTML 的半自动提取分析方法,也即:在信息提取之前通过解析器将 Web 文档解析成语法树,通过半自动的方式产生提取规则,将信息提取转换成对语法树的操作实现信息提取"。

在 Web 中,信息是以半结构化和无结构文档的形式组织存储的,参考文献[2]中指出:"这些数据没有统一的模式,数据的内容和表示相互交织,数据内容基本上没有语义信息进行描述,仅仅依靠 HTML 语法对数据进行描述"。

当前,对 Web 信息提取的研究主要有两个方向[3]:一种是研究怎样把网页中无结构或半结构化数据转换为结构化数据,这类研究的主要目的在于提取细粒度的数据;另一种研究则是希望通过信息提取技术,提取标题、正文等主题内容或兴趣区域。本文设计并实现的新闻中心平台,主要注重的是对新闻标题、链接、发表时间和正文等内容元素的提取。

由于现实中 Web 页面种类繁多,形式各异,在国内外的研究工作中,学者们提出了多种 Web 信息提取方式,例如:Finn[4]等人将 HTML 文档看作字符和标签组成的序列,在字符集中的区域提取文字。这种方式适用于以文字为主要内容的文档,而不便于提取文档中的图片、链接等内容;胡国平等[5]人针对新闻网站提取了基于统计的正文的抽取方法,但却只适合所有正文只有一个 TABLE 标签中的文档;而杨成[6]提出了一种面向由 XML 描述的 Web 文档的、基于用户主题信息的模式和数据提取方法。该方法利用学习算法从样本文档中提取规则,然后使用匹配算法从目标文档中训练出模式。

考虑到新闻领域相关的内容元素较为简单(包括标题、正文等),本文认为针对某个具体的新闻网站,可以由人工编码完成少量差异化的提取和分析工作,以插件形式加入到整个新闻提取和分析过程中,实现对各类新闻网站内容的准确高效提取。

## 2 新闻中心平台的设计

### 2.1 新闻生成器插件的设计

中间件是新闻中心平台的核心,是进行半自动化信息提取的部分。中间件主要类的 UML 类图如图 1 所示。图中,IGenerator 接口对提取新闻的主要方法(获得标题及正文等)进行了定义,凡是具有实现了 IGenerator 接口的类的插件均可以加入到平台中作为针对某个新闻网站的生成器,为平台提供来自某一特定网站的新闻。考虑到流行的新闻网站的架构具有一些相似之处,本新闻平台事先设计了一些实现 IGenerator 接口的基类,使用户在开发某些类型的新闻网站生成器插件时,无需从头开始编码工作,只需由这些基类派生,并实现一些差异化的工作即可。例如,TitleListNewsGenerator 基类对如图 2 所示类型的新闻网站提供了支持。

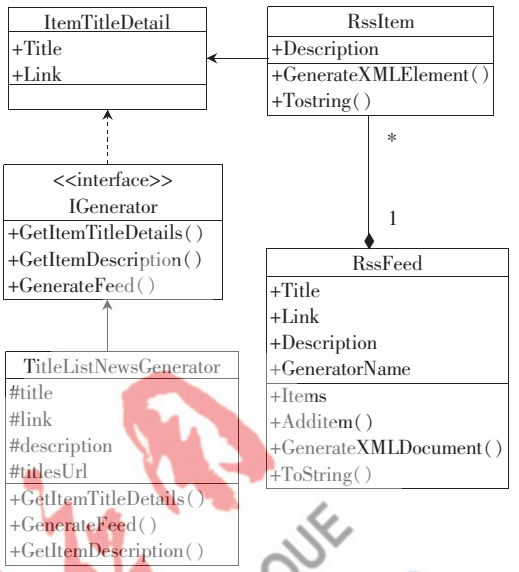


图 1 插件平台主要类的 UML 图类

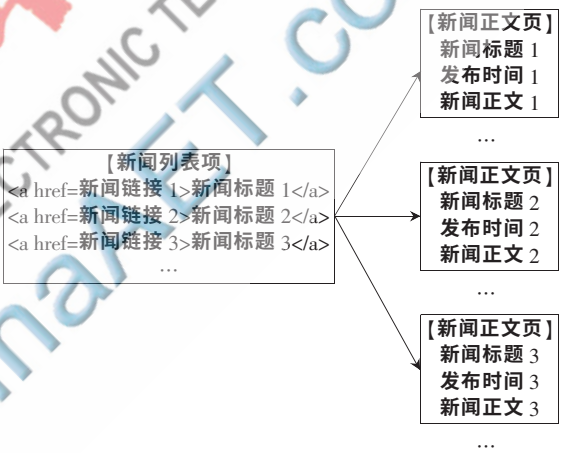


图 2 一种常见的新闻网站架构

TitleListNewsGenerator 基类对提取这类新闻网站信息的功能做了进一步的封装和优化(例如并行分析多个新闻正文页),如需从某个具体的、如图 2 所示的新闻网站提取信息,只需编写派生自 TitleListNewsGenerator 基类的实例,再对特定网页的少量特殊内容进行人工编码实现(例如若新闻链接地址是通过 JavaScript 事件生成的,则需要对这种链接进行转换)即可。提取这类新闻网站的信息的流程如图 3 所示。

针对不同类型的新闻网站,新闻生成器插件平台可以开发多种不同的基类,将其中可以自动化完成的工作预先实现,而插件制作人员只需对所针对的网站的特殊部分(例如与网页版面美化有关的少量特殊内容可能需要过滤)加以处理即可。

新闻生成器插件中包括至少一个直接或间接实现 IGenerator 接口的类,该接口定义新闻生成器必须实现的功能,包括获得新闻标题和新闻描述等。新闻生成器插

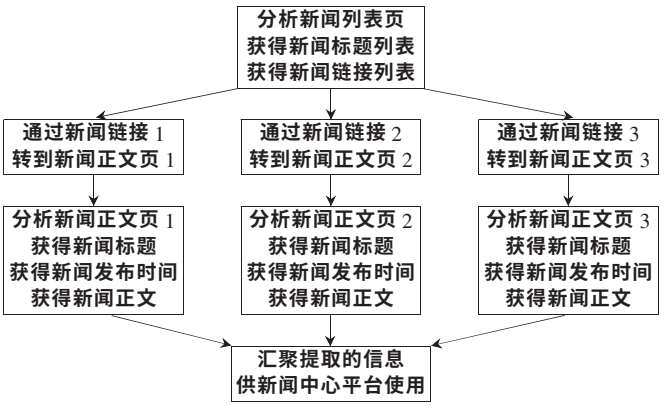


图3 一种常见的新闻提取流程

件针对不同的新闻网站，采取不同的分析页面源代码和抓取分析策略，最终实现在IGenerator接口中规定必须实现的功能。

新闻生成器插件在分析页面源代码的过程中，需要注意的是将页面图片、附件所使用的相对路径或由JavaScript在事件中生成的路径均转换为绝对路径，使其在从新闻网站独立出来后仍能正常使用。此外，在处理过程中若发生错误，应将异常交由日志记录模块进行记录，并由异常处理模块做相应处理。

## 2.2 新闻中心平台的设计

基于半自动化提取新闻的设计思想，为江苏城市职业学院实现了一套可扩展的新闻中心平台。其平台由归档数据库、新闻生成器插件平台、服务层、RSS种子生成模块、日志记录和异常管理模块以及多种类型的新闻中心表示层（客户端、网站、RIA等）组成，其整体架构如图4所示。

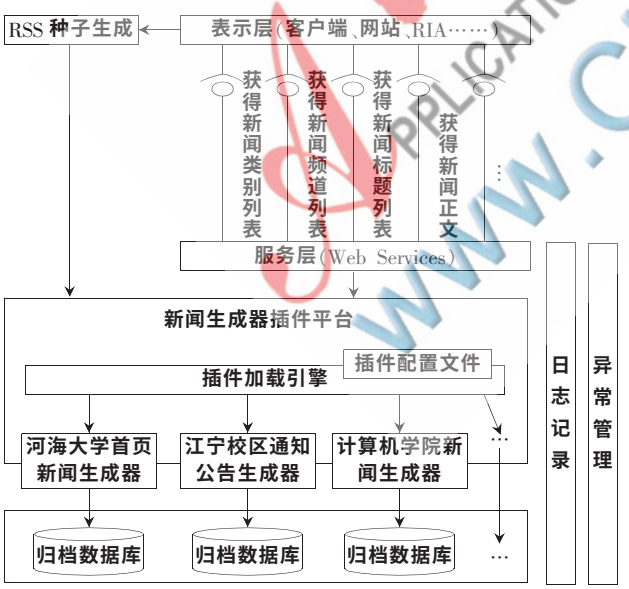


图4 新闻中心平台的架构

### 2.2.1 新闻生成器插件平台

基于2.1节插件的设计思想，新闻生成器插件平台负责将各插件集成到新闻中心平台中，为新闻中心平台提供来自不同网站的新闻信息。插件加载引擎通过读取插件配置文件定位插件并加载。插件配置文件是一个XML文件，该文件定义了各新闻生成器插件所属的新闻类别、插件文件路径和生成器类的完整路径，其格式如下所示：

```
<?xml version="1.0" encoding="utf-8"?>
 <categories>
   <category name="c1" title="t1" description="d1">
     <generator>
       <name>g1</name>
       <path>g1.dll</path>
       <class>G1</class>
     </generator>
     <generator>
     …
     </generator>
     …
   </category>
   <category name="c2" title="t2" description="d2">
   …
   </category>
     …
 </categories>
```

要为新闻中心平台增加新的新闻生成器插件，只需在插件配置文件中增加该插件的配置信息以及插件加载引擎即可定位到该插件，并通过调用相关方法，向新闻中心平台提供来自新来源的新闻信息。

### 2.2.2 归档数据库

归档数据库负责存放保留由各新闻生成器通过分析新闻所在网站源代码而抓取生成的新闻内容。

### 2.2.3 RSS种子生成模块

在新闻生成器插件平台的基础上，RSS种子生成模块能够为每个新闻来源生成一个RSS种子，以便用户使用RSS阅读器订阅新闻。

由于RSS文件是XML格式的，因此在生成RSS种子时，需要转换或过滤与XML文档不兼容的字符。另外，根据RSS标准的规范定义，需要将日期时间转换为RFC822规定的格式。

## 2.3 服务层

在新闻生成器插件平台的基础上，服务层进一步将功能抽象为一个个平台无关的Web Services方法，以适合为多种类型的表示层提供功能。

服务层主要提供以下服务：获得新闻类别列表、获得新闻频道列表、获得新闻标题列表和获得新闻正文等。

## 2.4 日志记录模块和异常管理模块

日志记录和异常管理模块贯穿整个新闻中心平台的服务周期，用以记录平台的工作状况，并在发生异常

时及时采取措施。

### 2.5 多种类型的新闻中心表示层

通过调用服务层提供的 Web Services，新闻中心的表示层可以使用不同技术，并设计成为多种不同的表现形式，从而满足不同用户的需要。

## 3 新闻中心平台的实现和运用

### 3.1 运行环境与实现

新闻中心平台基于.NET Framework 4.0 构建，除表示层因具体技术不同而有所区分外，新闻中心平台的其余部分最终均部署于 Dell PowerEdge R900 服务机上使用 VMWare ESX 3i 划分的一台安装有 Windows Server 2008 操作系统的虚拟机上。

新闻中心平台的客户端可以多种不同形式（网站、PC 或手机应用程序等）向用户提供新闻。图 5 是江苏城市职业学院新闻中心客户端的运行效果图。图中的新闻均由新闻中心平台通过提取与分析江苏城市职业学院网站的内容自动整理生成，并与网站更新保持同步，用户通过使用新闻中心平台，能够便捷地集中浏览原本散落于网站各个页面学院的新闻。

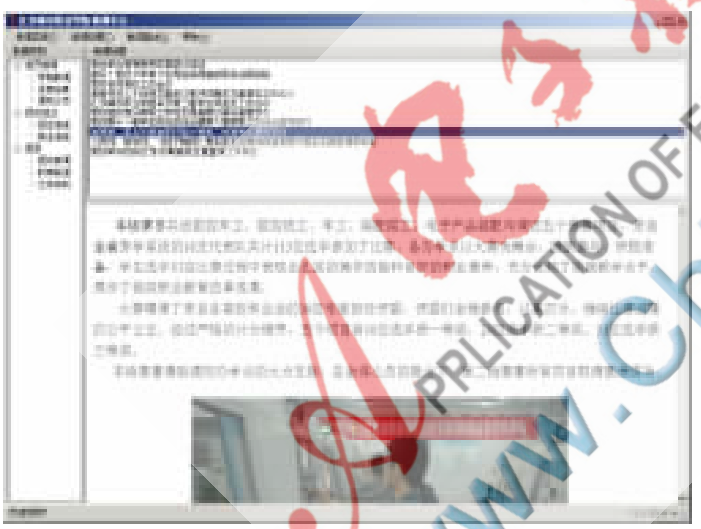

图 5 基于新闻中心平台的江苏职业学院新闻中心客户端运行截图

### 3.2 新闻中心平台的优势

除使用本文实现的新闻中心平台的方式外，还可通过设计门户网站或设计搜索引擎的方式方便对信息的获得，与此二种方式相比，新闻中心平台有其特有的优势。

#### 3.2.1 与门户网站相比的优势

新闻门户网站能够分类发布和整理新闻，但重新架构一套门户网站，意味着需要完全放弃现有的所有新闻网站，重新组织来自众多院系部门的新闻维护人员，成本极高。此外，各院系部门并非完全不需要一个展示自身的网站平台，完全废弃这些网站而统一使用门户网站将使各院系部门的形象千篇一律，流程僵化，而保留这些网站并同时使用门户网站又将造成信息的冗余。

而新闻中心平台是利用插件半自动化地从各院系部门现有网站中提取新闻，避免了重复建设，并只由归档数据库做适当缓存，随时可以删除，不存在大量冗余。与建设门户网站相比，使用新闻中心平台能有效降低成本投入，且更加灵活。

#### 3.2.2 与搜索引擎相比的优势

可以使用搜索引擎，通过关键词检索新闻。然而检索的范围和粒度都难以控制，也不能满足一般的新闻浏览需要。而新闻中心平台将新闻分门别类，在方便统一浏览的同时，可以足够快速地找到所需的新闻，在必要的时候还可以加入搜索功能作为辅助。

本文设计并实现了一套基于插件架构的新闻中心平台，该平台在新闻中心网站和客户端的设计中进行了实践，通过插件提取新闻网站的内容进行分析汇总，解决了新闻来源混乱而不易于获得的问题，且具有较好的扩展性。今后将进一步分析新闻网站的特点，有针对性地提高信息提取的自动化程度。

参考文献

[1] 王宇宁.隐马尔可夫模型在信息抽取中的应用研究[D].大连：大连理工学院，2007.

[2] 袁宇丽. 基于 HTML 网页的 Web 信息提取研究 [D].成都：电子科技大学，2005.

[3] 谢德辉. 面向刑侦网页的信息抽取与主题爬虫应用研究[D].大连：大连理工学院，2007.

[4] FINN A, KUSHMERICK A, SMYTH B. Fact or fiction: content classification for digital libraries [C]. The 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland, 2001: 110-115

[5] 胡国平，张巍，王仁华.基于双层决策的新闻网页正文精确抽取[J].中文信息学报，2006，20(6)：1-10.

[6] 杨成.基于 XML 的网页信息提取系统的研究与设计[J].电脑知识与技术，2009，5(25)：7327-7329.

（收稿日期：2012-09-01）

作者简介：

李剑，男，1975 年生，硕士，工程师，主要研究方向：软件工程。

陈海建，男，1976 年生，博士，副教授，主要研究方向：计算机软件与电子商务。