

# 基于模板的 Web 信息提取系统的设计与实现

周合明, 奚建清

(华南理工大学软件学院, 广东 广州 510006)

**摘 要:**随着 Web 上信息的迅速扩展, Web 信息提取技术正应用于搜索引擎、用户兴趣挖掘以及个性化信息获取等多种应用和研究中。文中通过采用模板技术, 设计并实现 Web 招聘信息提取系统。该系统根据已配置的模板, 从各高校就业中心网站提取招聘信息, 结构化并存入内嵌 HSQL 数据库中。基于已采集的信息, 进行个性化搜索, 找到用户所需招聘信息。试验结果表明, 该系统能够完成信息的提取和个性化搜索, 具有很好的实际效用。另外, 由于该系统采用 java 和内嵌 HSQLDB 开发, 具有高度的平台移植性和很好的移动便捷性。

**关键词:**信息提取; 模板; 内嵌 HSQL; 个性化搜索

**中图分类号:** TP31

**文献标识码:** A

**文章编号:** 1673-629X(2011)11-0105-04

## Design and Realization of Template-Based Web Crawler

ZHOU He-ming, XI Jian-qing

(School of Software Engineering, South China University of Technology, Guangzhou 510006, China)

**Abstract:** With information rapidly expanding in the Web, extracting information for Web page is applying many fields. Design and realize a Web Crawler by using template technology. This Crawler extracts information from employment online of college through configured templates, structure information and store in the In-Process HSQLDB. Based on information stored in-database, perform customized search and find useful recruitment informations for user. Experimental results show that this system is able to complete the extraction of recruitment information and customized search. This system has high practicability. In addition, the system is developed by java and In-Process HSQLDB, so has the high platform portability and very good convenience.

**Key words:** extracting information; template; In-Process HSQL; customized search

## 0 引言

据统计, 2011 年全国大学毕业生数量将达 650 万人, 毕业生首先面临的是找工作问题, “就业难”已经成为毕业生心头的一块大石, 随着高校毕业生的持续增长, 在未来相当长的时间内, 大学生就业压力不会减弱。如何才能在就业大军中脱颖而出, 率先找到一份合适的工作是毕业生最关心的。为了找到一份合适的工作, 除了提升个人能力之外, 还得拓展信息源, 及时获得所需求的招聘信息。目前毕业生主要通过以下渠道获取招聘信息: 第一, 学校就业指导中心网站, 这个主要信息来源, 也更适合毕业生, 但不是每个学校都有就业指导中心网站, 另外一个学校的就业指导中心所发布的招聘信息毕竟有限, 不能满足毕业生的信

息需求; 第二, 专业招聘网站, 这也是毕业生获得招聘信息的主要来源, 但该类网站不只是面向应届毕业生, 还面向社会人士, 需要工作经验, 大部分招聘信息对应届毕业生不合适; 第三, 通过搜索引擎获取招聘信息, 通过这种方式也能获得一些招聘信息, 但搜索引擎追求普适的设计目标, 不能满足个性化信息检索, 获取的信息也不及时。

通过以上比较, 比较符合毕业生的信息源仍旧是学校就业指导中心网站, 为了克服单个学校就业指导中心信息量不足的问题, 文中设计并实现了一种基于模板的 Web 信息提取系统, 本系统将根据相应模板从各大高校就业指导中心提取招聘信息并存入数据库中, 然后对已采集的信息进行个性化搜索, 满足个性化信息需求。

收稿日期: 2011-04-18; 修回日期: 2011-07-22

基金项目: 广东省产学研结合项目(2009A09010035); 广东省国际合作项目(2009B050700008); 广东省科技工业攻关项目(2008B09050019, 2006B80407001)

作者简介: 周合明(1986-), 男, 硕士研究生, 研究方向为软件工程; 奚建清, 博士生导师, 研究方向为数据库系统。

## 1 国内外研究现状

Web 信息抽取就是从 Web 页面所包含的无结构、半结构或者结构化的信息中识别用户感兴趣的数据, 并将其转化为结构和语义更为清晰的格式(XML、关系数据等)<sup>[1]</sup>。目前国内外已经有许多 Web 信息提

取方法,其主要有两类:一是基于模板<sup>[2-5]</sup>,通过为信息源建立相应的模板库,将 Web 页面文档与模板库中的模板进行匹配而实现,模板的表达能力直接影响提取结果的准确性;二是基于本体(Onology)<sup>[6-8]</sup>,通过构建领域本体,把本体解析为一系列的概念和关系,然后利用这些概念与关系对文本进行语义分析,完成信息提取,其关键在于构造良好的本体库。

目前本体的相关研究日益成熟,但是运用本体来解决实际的问题有一个直接的瓶颈—本体库。本体库的构建一般使用半自动化的方式,要求相关领域专家和计算机专家进行有效的协作开发,这在实际项目中对开发人员的要求高。

本系统将采用模板进行 Web 信息提取,构建模板的方法有两种:一种是采用页面结构特征分析、归纳学习等方法自动获取页面模板,如文献[2~5],另一种就是采用手动配置页面模板,如文献[9]。本系统将采用手动的方式配置模板。

使用此方案主要从以下四个方面考虑:

- (1) 通过手动配置的模板结构丰富,表达能力强;
- (2) 信息提取速度快、质量高,信息详细度高,有利于个性化搜索;
- (3) 由于信息采集的信息源固定,就是各大高校的就业中心网站,可以很容易地制定信息源模板;
- (4) 信息源变动可能性小,不像商业网站,各大高校的就业中心网站一般只是简单地发布招聘信息,变动性小,这样对某个就业中心网站制定的模板就可以达到一次制定,持续使用,无需更改。

## 2 系统介绍

### 2.1 系统结构

如图 1 所示,本系统主要由三部分组成:信息采集器、信息管理模块、UI。信息采集器根据模板采集信息并结构化注入 Hibernate 对象中,信息管理模块完成信息的持久化和个性化搜索服务,UI 为用户提供友好的操作界面,方便个性化搜索设置、搜索结果展示以及系统设置等操作。

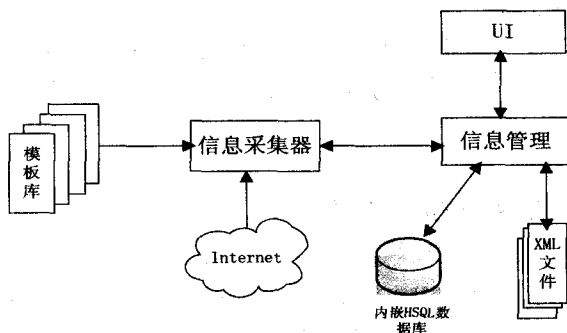


图 1 系统总体结构图

### 2.2 信息采集器

信息采集器是本系统的核心模块,完成 Web 页面信息的提取并进行结构化,其结构如图 2 所示。它主要由四个子模块组成:采集控制器、连接器、信息提取器和信息组织子模块。

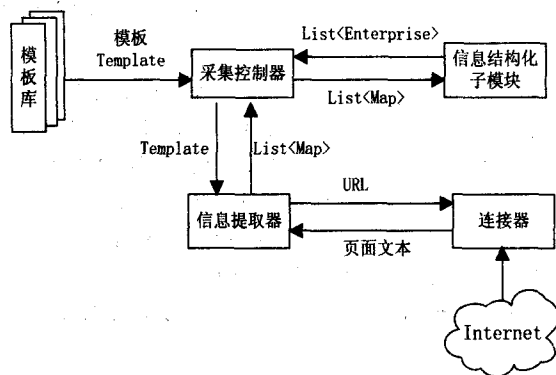


图 2 信息采集器结构图

#### 2.2.1 采集控制器

采集控制器是信息采集器的中央控制器,由它来协调各个子模块,完成信息的采集并结构化。采集控制器依次读取模板库中的模板文件,然后使用信息提取器,提取模板中指定的信息源信息,最后使用信息结构化子模块完成信息的结构化。

#### 2.2.2 信息提取器

信息提取器根据模板中制定的提取规则对获得的页面文本进行信息提取。信息提取结果保存在一个 List 列表中,列表的元素为 Map 对象,采用键/值对的方式保存匹配结果,例如使用一个 Map 对象保存一家企业信息。模板结构和提取算法将在第三节介绍。

#### 2.2.3 连接器

连接器根据相应的 URL,获取页面文本,如果需要登陆认证,则需要传递认证参数先进行认证,通过认证后再获取目的页面文本。连接器主要使用 HttpClient<sup>[10]</sup>包中的 HttpClient、PostMethod、GetMethod、NameValuePair 类进行 get 连接和 post 连接,获取页面文本。

#### 2.2.4 信息组织子模块

信息组织子模块接收页面文本提取结果并进行结构化,使用 java 反射机制,把提取的招聘信息注入企业实体(Enterprise)对象中,这有利于信息管理和信息持久化。

### 2.3 信息管理模块

信息管理模块通过信息采集器采集招聘信息,对已经结构化的招聘信息进行管理,完成信息的持久化和信息的个性化搜索。信息的搜索与持久化由 Hibernate 框架实现。

### 2.4 UI

UI 为用户提供友好的操作界面,使用 swing

包完成开发。用户可以通过 UI 手动启动信息采集器进行信息采集,也可以通过设置环境变量定时进行信息采集。通过设置搜索条件完成对采集结果的搜索,满足个性化信息需求,搜索结果将被分页展示。

### 3 关键技术

#### 3.1 模板

本系统采用模板技术提取 Web 页面信息,把 Web 页面看作特殊字符串,根据模板中指定的信息提取规则,使用字符串处理技术提取 Web 页面信息。文献[9]中也使用了类似模板技术,但该文献中的模板存在以下几点不足:

(1)不支持多级模板的配置,不支持多级页面抓取;

(2)模板文件采用简单的 key=value 形式,表达能力不够;

(3)模板结构单一,扩展性不好。

为了克服以上几个不足,本系统使用 xml 格式配置模板,提供了丰富的配置字段,具有很强的表达能力和很好的扩展性。

配置模板的根元素为<Template>,由五个子元素构成:<Page>、<Type>、<Rules>、<TemplateHandlers>、<Source>。其中<Page>、<Rules>是必须的,其它可选,这些元素最多可以出现一次。

<Page>元素指定了页面 URL 生成策略以及连接参数,通过其 class 属性配置相应类,所配置的类必须实现 PageStrategy 接口,此接口包含三个方法:nextURL()用于获取下一个页面 URL,getMethod()用于获取发送请求的方法,getParameters()用于获取请求参数。

本系统已经实现两个 URL 生成策略类:PageIdStrategy、URLCollectionsStrategy,其中 PageIdStrategy 根据页面翻页关键字依次生成页面 URL,URLCollectionsStrategy 根据配置指定的页面 URL 集获取页面 URL。

<Type>元素指定本模板提取结果类型,本系统只有两种类型:企业信息类 Enterprise 和职位信息类 JOB。

<Rules>元素指定信息提取规则集,其主要有 2 个子元素构成:<Rule>、<Rule-

Handlers>,其中<Rule>是必须的,至少出现一次,<RuleHandlers>是可选的,最多出现一次。

<Rule>元素指定一条提取规则,其包含以下 8 种子元素:

(1)<BeninTag>:开始标识符。

(2)<EndTag>:结束标识符。

(3)<Name>:属性名称。

(4)<Editor>:提取结果编辑器,该编辑器必须实现 Editor 接口。

(5)<DefaultValue>:默认值。

(6)<ChildRules>:子规则集。

(7)<ChildTemplate>:子模板。

前面三个为必须子元素,这些元素最多可以出现一次。

<RuleHandlers>元素指定了规则级别的处理器集合,用于处理信息提取结果 Map,其包含一个或多个子元素 RuleHandler 子元素,通过 RuleHandler 元素的 class 属性指定一个处理器类,该类必须实现接口 RuleHandler 接口,该接口只包含一个方法 handle (Map);返回 true 将保存提取结果 Map,返回 false 将丢弃提取结果 Map。

<TemplateHandlers>元素指定了模板级别的处理器集合,用于处理信息提取结果列表 List<Map>,其包含一个或多个 TemplateHandler 子元素,通过 TemplateHandler 的 class 属性指定一个处理器类,该类必须实现 TemplateHandler 接口,该接口包含一个方法 handle (List<Map>),无返回值。

<Source>元素指定信息来源。

图 3 为华南理工大学就业在线模板样例。

```
<Template>
  <Page class="org.scut.edu.urlstrategy.impl.PageIdStrategy">
    <Property key="url">http://202.38.193.238:8080/information/index.jsp</Property>
    <Property key="method">get</Property>
    <Property key="pageKey">pageNO</Property>
    <Property key="beninNumber">1</Property>
    <Property key="endNumber">2</Property>
  </Page>
  <Rules>
    <Rule>
      <BeninTag><CDATA[[<a href="companyDetail.jsp?"]</BeninTag>
      <EndTag><CDATA[["</a>]]</EndTag>
      <Name>companyName</Name>
      <Edit class="org.scut.edu.edit.impl.DateEditor">
        <Property key="dateRegex">(\d)+-(\d)+-(\d)</Property>
        <Property key="datePattern">yyyy-MM-dd</Property>
      </Edit>
      <ChildTemplate>
        <ConfLocation>scutcompany.xml</ConfLocation>
      </ChildTemplate>
    </Rule>
    <RuleHandlers>
      <RuleHandler class="org.scut.edu.handler.impl.FilterByDateHandler">
        <Property name="day">1</Property>
      </RuleHandler>
    </RuleHandlers>
  </Rules>
  <SourceName>华南理工大学就业在线</SourceName>
</Template>
```

图 3 华南理工大学就业在线模板

### 3.2 提取算法

根据模板 Template 中的规则集 Rules 对页面文本进行循环模式匹配,对匹配的结果进行处理并保存在 Map 对象中,所有的提取结果 Map 对象保存在列表 List 中,其主要流程如下:

(1) 读取模板 Template 中配置的规则集 Rules,创建一个 Map 对象 map,针对规则集中的每个规则 Rule 执行以下子过程:

① 根据 Rule 中配置的 BeginTag 和 EndTag 对页面文本进行模式匹配,提取一条匹配文本 m,如果模式匹配失败, isOver=true,跳出循环。

② 如果 Rule 中配置了子模板,将递归调用信息提取过程,把提取结果 result 放入 map 中,此时的 key=Rule. Name + “. ChildTemplate”,即 map. put ( key, result)。

③ 如果 Rule 中配置了子规则,将用子规则对文本 m 进行信息提取,并把结果放入 map 中,此时的 key=Rule. Name+ “. ChildRules”。

④ 使用一组过滤器对文本 m 中的 html 代码进行过滤和替换。

⑤ 调用 Rule 中配置的编辑器 Editor 对过滤后的文本 m 进行处理,把处理返回的结果存入 map 中,此时的 key=Rule. Name。

(2) 依次调用 Rules 中配置的处理程序 RulesHandler 对 map 进行处理,如果返回 true,将把 map 放入列表 List 中,如果返回 false 将不保存,如果捕获到异常将置 isOver=true。

(3) if(isOver) 执行步骤(4),否则跳到步骤(1)继续提取。

(4) 依次调用 Template 中配置的处理程序 Template-Handler 对提取结果集 List 进行最后处理。

(5) 提取结束,返回提取结果列表 List。

### 3.3 持久化技术

数据持久化就是将内存中的数据保存到磁盘加以“固化”。为了满足个性化搜索,必须把信息采集结果进行持久化。本系统使用 Hibernate<sup>[11]</sup> 框架实现采集结果持久化。Hibernate 是一个开源的 ORM 框架,使用面向对象的编程思维来操作关系数据库,完成对象与关系之间的映射。

从系统的性能和便捷性方面考虑,本系统使用的是 HSQL<sup>[12]</sup> 数据库, HSQL 数据库是一款纯 Java 编写的开源数据库,体积小,不到 1M,最重要的一点是 HSQL 可以在 In-Process 模式下运行,在该模式下,数据库引擎作为系统的一部分在同一个 JVM 中运行,这样不仅访问速度快,而且很容易使数据库随系统一起移植。

## 4 试验结果

使用已经实现的系统对国内 15 所高校的就业指导中心网站最近一个月发布的招聘信息进行模拟提取,并进行个性化搜索,其中搜索条件:求职类别为全职,发布日期为一个月和职位名称关键字为 java,结果如图 4 所示,共搜到 27 条记录,每页显示 10 条,双击职位名称可以查看职位详细信息,双击公司名称可以查看公司详细信息,双击信息来源将打开信息来源页面。试验结果表明,本系统可以完成招聘信息的提取以及提供个性化搜索。

File Edit Run Option

工作地区： 求职类别： 全职

发布日期： 最近一个月 职位名称： java

序号	职位名称	工作地点	求职类别	招聘人数	发布日期	信息来源	
1	Java开...	广州...	全职	广州...	2011-04-14	华南理工...	
2	Java产...	上海...	全职	暂无	5人	2011-04-14	上海理工...
3	Java网...	上海...	全职	暂无	5人	2011-04-14	上海理工...
4	游戏开...	掌上...	全职	暂无	暂无	2011-04-13	北京科技...
5	数据库...	博望...	全职	暂无	若干	2011-04-12	上海理工...
6	JAVA技...	上海...	全职	暂无	若干	2011-04-12	上海理工...
7	java工...	中视...	全职	暂无	若干	2011-04-12	北京科技...
8	Java开...	广东...	全职	广州	2	2011-04-12	华南理工...
9	高级游...	广州...	全职	暂无	若干	2011-04-12	华南理工...
10	Java开...	广州...	全职	天河...	2	2011-04-12	华南理工...

共27条记录-每页10条-第1页

上一步 下一步

图 4 职位搜索结果

## 5 结束语

本系统通过对各高校就业指导中心网招聘信息的采集,解决了单一学校信息不足的问题,个性化的搜索满足了毕业生个性化信息需求。另外,文中改进的基于模板的 Web 信息提取方法不仅在性能、准确率都优于传统方法,准确率接近 100%,而且由于增加了抓取结果编辑器和不同级别的处理器等新特性,可以在不修改源代码的情况下,很容易的在 Web 信息提取过程添加新的定制行为,具有很强的表达能力和扩展性,对于像高校就业指导中心这种规模不大的数据源,具有很好的应用效果。本方法也存在一定的局限性,由于模板需要手动制定,因此对制定模板的人员有一定的计算机基础要求,另外如果需要拓展数据源,手动制定模板比较费时,出错率也增大,版本库的维护也会随之困难起来。未来可以通过手动与自动相结合的方式制定模板,提高效率,降低出错率。

### 参考文献:

- [1] 刘 辉,陈静玉,徐学洲. 基于模板流程配置的 WEB 信息提取[J]. 计算机工程,2008,34(27):55-57.
- [2] 郑长松,傅 彦,余 莉. 基于模板的 Web 信息自动提取方法[J]. 计算机应用研究,2005,26(2):570-572.
- [3] 郭太飞,何洁月. 归纳学习 XPath WEB 信息提取规则[J]. 计算机技术与发展,2007,17(3):98-101.

(下转第 112 页)

```
hsqldb;file:/opt/sslexplorer/db/explorer_configuration",
"sa", "" );
```

```
Statement stmt = conn.createStatement();
```

```
//打印 users 表里面的内容
```

```
//插入一条记录 每次执行这个语句的时候请更
改用户名,不然会插入相同的用户名
```

```
//ENCPASSWORD 内置函数用来加密密码
```

```
//curdate()
```

做完以上工作,不管用户是在注册页面注册还是管理员添加,数据显示是同步的。

### 3.2 帐户的建立和配置

#### (1) 建立组和帐户。

以管理员的身份登陆后,可建立一些组和帐户,该建立过程较简单,根据提示即可完成。在本例中建立一个“信控学院”的组,并在该组下建立一个帐户“小王”。

#### (2) 策略的建立和配置。

因刚建立的帐户是不活动的,必须对其分配一个策略,才能变为活动的,该帐户才能使用。其操作步骤是在管理控制台 (Management Console) 中选择访问控制 (Access Control) 中的策略,然后选择创建策略即可根据提示步骤完成该策略的建立和配置。完成后,换到帐户的界面,即可看到刚建立的帐户“小王”的状态变为活动。

### 3.3 资源管理分配

在建立好具体的帐户和策略后,便可对这些帐户分配资源。在资源管理分配中主要有帐户登陆 the SSL Explorer VPN server 方式的建立、Web Forwards 的建立、Network Places 的设置等。对于帐户登陆方式的建立主要是使所设置的帐户能远程登陆并可控制桌面;Web Forwards 的建立主要是因校园网中很多的软件和应用系统都是基于 Web 服务的,如 OA 系统等,该部分的设置非常重要;Network Places 的设置主要是为了解决局域网中的一些资源的访问、网上邻居、FTP 系统的安全访问。以上资源管理的设置分配的具体步骤

可参照软件中的帮助文档。

## 4 结束语

利用 SSL Explorer 构建校园远程访问的 VPN 具有灵活性、安全性、经济性和扩展性等优点,非常适合校园的远程访问。随着高校师生对电子资源的利用率越来越高,特别是许多科研工作及学习的地理范围早已超越了校园的限制,因此构建校园远程访问的 VPN 是必然的要求和发展的趋势。

### 参考文献:

- [1] 李 玮,侯整风. SSL 协议安全缺陷分析[J]. 计算机技术与发展,2006,16 (12):224-226.
- [2] 牛少彰,郭延玲. SSL VPN 原理及其优势[J]. 通信市场,2006(10):51-52.
- [3] 陈小中. SSL VPN 技术与校园信息化[J]. 科技资讯,2008 (34):26-26.
- [4] Zhang Zhensheng, Zhang Yaqin, Chu Xiaowen, et al. An Overview of Virtual Private Network (VPN): IP VPN and Optical VPN[J]. Photonic Network Communication, 2004, 7(3):213-225.
- [5] 郭 玲,李伟生. SSL VPN 的设计与实现[J]. 计算机技术与发展,2007,17(8):148-150.
- [6] Broderick J S. VPN Security Policy [R]. [s. l.]:[s. n.], 2001.
- [7] 曾巧红,徐文贤,林绮屏. 基于 SSL VPN 的图书馆远程访问系统的构建[J]. 情报科学,2007 (10):1520-1524.
- [8] 李 毅,黄晨晖. 基于 SSL VPN 实现远程快速安全访问校园网[J]. 中国教育信息化,2007(3):25-26.
- [9] 马军锋. SSL VPN 技术原理及其应用[J]. 电信网技术,2005(8):6-8.
- [10] 徐 忻. 利用开源软件实现基于 SSL VPN 的图书馆访问[J]. 现代情报,2009,29(4):160-163.
- [11] Heller P. Remote access: its impact on a college library[J]. The Electronic Library, 1992, 10(5):287-289.
- [12] 廖继业,李双玲. 图书馆电子资源提供校外访问安全问题研究[J]. 新疆教育学院学报,2008(4):147-149.

(上接第 108 页)

- [4] Liu Yaqing, Chen Rong, Yang Hong. Web Information Extraction Based on Hierarchical Model [C]//The 2009 International Conference on Computational Intelligence and Software Engineering. [s. l.]:[s. n.], 2009:1-5.
- [5] Ma Jun, Li Tihong. XML-based Web information extraction system design and implementation [C]//The 2010 3rd IEEE International Conference on Computer Science and Information Technology. [s. l.]:[s. n.], 2010:551-554.
- [6] 王继东,张 瑜,李 娜. 基于本体的语义检索技术研究 with 实现[J]. 计算机技术与发展,2009,19(10):134-137.
- [7] 郭 猛,冯志勇. 基于本体实现有效 Web 信息检索[J]. 微处理机,2007(4):116-119.
- [8] 拜战胜,徐德智,彭佳红,等. 基于主题本体的信息采集模型研究[J]. 计算机技术与发展,2009,19(10):102-105.
- [9] 王晓地,奚建清. 网上信息定制搜索技术的研究[J]. 江南大学学报(自然科学版),2010(9):45-49.
- [10] HttpClient [EB/OL]. 2010. <http://hc.apache.org/httpcomponents-client-ga/index.html>.
- [11] Hibernate [EB/OL]. 2009. <http://www.hibernate.org/>.
- [12] HSQldb [EB/OL]. 2010. <http://hsqldb.org/>.