网络店铺信息自动提取

郑思婷[1], 杨烜会[2], 王周敬[1]
(1.厦门大学自动化系 福建 厦门 361005 2.厦门理工学院商学系 福建 厦门 361024)

【摘 要】: 以国内 C2C 行业最具代表性的交易平台--淘宝网为例,设计了针对网络店铺的信息自动提取流程,并利用 Python 语言实现了对网络店铺信息的自动采集和结构化输出。

【关键词】: 信息提取; 网络店铺; 正则表达式; Python

0.引言

随着电子商务的蓬勃发展,网络购物日渐兴旺,网络店铺(以下简称网店)页面中所包含的店铺信息、商品信息,以及服务信息也不断积累,形成了巨大的信息资源库。信息提取(information extraction)--利用计算机自动从网店页面中收集有用信息, 是利用该信息资源库的必由之路,是商业智能和管理研究的数据源头。

为方便从海量的网店页面中快速、准确地获取信息,本文设计了流水线式提取流程,并用 Python 语言实现了自动提取工具。

1、网店页面的特点

网店页面具有半结构化特征。

一方面,网店页面包罗万象。网店页面本身是大量文本、图片或其它多媒体字符流的集合,因此无法通过结构化查询语句来处理网页信息[1]。同时,网店页面中还包含大量广告、导航等与商业主题不相关的信息[2],这些信息会干扰理解网店页面中所包含的商业意义。

另一方面,网店页面有一些规律可用于信息提取。网店页面和其它 Web 网页一样, 都用 HTML 书写而成,遵从 W3C 规范。其次,C2C 网店平台要求网店套用平台模板,网店页面要遵循平台提供者(如淘宝、拍拍等)规定的框架规则,这些规则也可为信息提取提供启发线索。

2、网店信息自动提取的设计与实现

基于网店页面的半结构化特征, 本文设计了一个网店信息自动提取流程,先过滤掉无关信息,然后利用特定标记匹配有用的网店信息。

2.1 网店信息提取流程

我们设计的网店信息自动提取流程如图 1 所示。首先,利用网络爬虫(web crawler)获取并保存网店页面到 HTML 文件;其次,通过过滤器去掉 HTML 格式标记和无关信息,形成页面信息的文本文件;最后,利用页面信息文件中语义标记来匹配和提取网店信息到特定的数据结构,进行 CSV 格式化输出。该提取流程采用流水线(pipeline)架构,以文本文件为处理媒介。

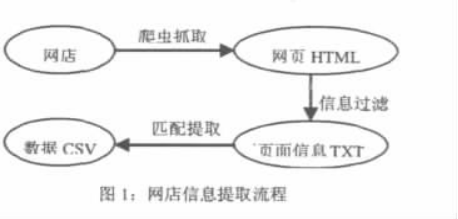

图 1: 网店信息提取流程

网络爬虫抓取是通用技术,有多种成熟的实现,本文不再赘述。下文将结合淘宝网店页面规范,重点介绍网店信息提取特有的技术细节--信息过滤和匹配提取。

2.2 HTML 文件的信息过滤

通用的信息提取技术有很多[3],但针对网店信息提取的目标和网店页面的特点, 我们选用 HTML 标记和关键信息的正则表达式过滤方法。

正则表达式通过分析各种信息元素所特有的呈现方式,构造字符串模式,在文本中检索或替换符合该模式的文本, 它是从非结构化文本中提取有用信息的利器[4]。淘宝网使用独立的 CSS 文件将页面渲染格式剥离出来,而用 HTML 主文件存放店铺信息、商品信息和服务评价信息。所以,我们可以用正则表达式将 HTML 主文件中的格式控制、HTML 标记等无关信息过滤掉。现代的程序设计语言,大都配备了正则表达处理模块,这里选用灵活的动态语言 Python 作为实现工具[5]。HTML 文件过滤的关键程序如下:

```
① for line in read.readlines():
② line=re.sub('<[^>]+>','',line)
③ line=re.sub('&#.+;','',line)
④ line=re.sub('&.+;','',line)
⑤ line=re.sub('+|\t+','',line)
⑥ if line[:-1].strip():
⑦ write.writelines(line)
```

图 2:HTML 文件过滤代码

图 2 中的代码逐行读入 HTML 主文件 (行①),然后运用 Python 正则表达式处理模块 re 中的替换功能,将 HTML 标签中的内容 (行②), 以及特殊字符串、空

本文得到厦门理工学院高层次人才项目(YSK09004R)资助

格、制表符和空行(行③④⑤⑥)等噪声替换为空字符串进行删除；最后将剩下的内容写入页面信息文本文件中(行⑦),从而完成 HTML 文件过滤。

在过滤 HTML 文件的过程中，要对图片所包含的信息进行特殊处理。例如,淘宝店铺中商家的信用信息是用淘宝专有的信用等级图标(红心、蓝钻、蓝冠、黄冠)来表示的。在过滤 HTML 文件的时候,需要提前注意到这些图形化信息，并使用特定字符串来替换这些图片,才能顺利地实现后续的信息提取。

2.3 页面信息的匹配提取

以过滤后的信息文本文件为基础，我们可以使用索引术语技术来进行网店信息提取[6]。

索引术语技术假定文本库中的文件语义和用户的信息需求语义可用同一组索引术语来表示[7]。淘宝网店页面有相对固定的格式,例如,累计信用的信息文本片段为" 累计信用:s_xxxx_n", 其中"s_" 是固定前缀,"xxxx"为信用类型,"n"为信用等级。要提取累计信用信息,以"累计信用"为索引即可满足索引术语技术的前提假设。在信息文本文件中,店名、累计信用、好评率、卖家服务态度、宝贝数量等重要的店铺信息都可以找到对应的索引标记,我们可以按图索骥,利用这些索引来标识并提取有用信息。

我们仍以提取累计信用信息为例来说明索引术语技术的工作原理。假设信息文本文件中的字符串为"累计信用:s_blue_3",通过分析淘宝网的信用等级信息可知,以上信息中需提取的信息是"blue"和"3"。信息提取的代码如下：

```
①  a=re.compile(u´累计信用:\w_\w+_\w´)
②  if a.search(line):
③  p=a.search(line).group().split(u"_")[1]
④  q=a.search(line).group().split(u"_")[2]
```

图 3:信息提取示例代码

首先,编译累计信用信息索引术语的正则表达式"累计信用:\w_\w+_\w"(行①);然后,在信息文本文件中搜索匹配(行②)。若该索引不存在,则进行下一个信息匹配。若存在,则从"累计信用:s_blue_3"文本串中提取具体的信用等级信息"blue"和"3"(行③④)。最后,通过对照事先建立的信用等级字典 rank=(´red´:u´ 红心 ´,´blue´:´ 蓝钻 ´,´cap´:´ 蓝冠 ´,´crown´:u´ 黄冠 ´),就可以给出具体的信用等级。例如,"s_blue_3" 表示"3 级蓝钻"。

利用索引术语技术,通过类似的方法,我们可以逐一提取信息文本文件中的有用信息，如店名、累计信用、好评率、卖家服务态度、宝贝数量等,然后写入特定的数据结构中。

由于各种主流的关系数据库管理系统 (如 Oracle, mySql)和数据处理软件(如 Excel 和 SAS)都支持纯文本的 CSV 数据格式,所以我们选择 CSV 作为数据存储格式,以方便后续的数据处理。Python 有专门的 csv 模块支持 CSV 格式读写，稍加改造以支持中文编码,就可进行格式化输出。

3、网店信息提取实例

图 4 的实例说明了提取淘宝网某店铺信息的过程。该网店页面的局部截图如图 4 (a)。网店页面的 HTML 文件片段如图 4(b)。可以看出,除了有用信息,该 HTML 片段还包括了大量的噪声字符串。HTML 文件经过过滤处理,得到如图 4(c)所示的信息文本文件。对信息文本文件进行匹配提取,并按 CSV 格式进行格式化,就可以输出如图 4(d)所示的结构化信息。

(a) 网店局部截图　　(c) 页面信息片段

描述相符：4.5低于4.25%
服务态度：4.8高于44.71%
发货速度：4.6低于2.31% <h4>服务</h4>
<img src="http://img.taobaocdn.com/tfb_person_small.gif?1235" border="0" align="absmiddle" alt="支付宝实名认证" title="支付宝实名认证"> <h4>店铺信息</h4><ul class="shop-attach"><li class="goodrate">好评率：99.67%
宝贝数量：175
<li class="setuptime">创店时间：2008-09-08

(b) 页面 HTML 片段

(d) 数据 CSV 片段

图 4: 某网店信息提取实例

4、结语

上面介绍了包括爬虫抓取、信息过滤和匹配提取在内的信息提取流水线过程，淘宝网店的实例也充分展示了该流程的有效性。将上述的信息提取工具稍加修改,即可用于其他的网店平台(例如拍拍、eBay 等)的店铺信息提取。

本文所述的技术能够自动搜集网店的运营信息,在商业上可用来监控网购品类，也可用来收集和跟踪竞争对手的运营情况。此外,也可以用类似的技术开发购物代理机器人(Shopping bot),在多个平台的多个网店中搜集和比对商品的价格、服务和反馈信息,为网购客户提供购物建议。发展网店信息提取技术,开发网店信息库能让网络购物更加便捷,网店竞争更加充分,具有重要的经济和社会价值。

(下转第 37 页)

受到实际工作中对计算机能力的要求，充分激发学生的学习动力。例如,针对文档编辑与排版模块,设计了"书信制作"、"制作课表"、"报纸排版"、"制作奖状"和"毕业论文排版"5 个任务，将使用 Word 工具的能力从基础到综合融入到这 5 个任务中。任务由浅入深,注重知识的连贯性,前面的任务内容为后面的课程打下基础。后面的任务在不断注入新知识和新概念的同时 ,也不断加入已学知识的复习内容，使学生在反复不断的学习过程中增强计算机技能。总之,教师在设计任务时应注意将各项理论知识和实践应用分解到每个任务中,突出教学重难点,明确教学目标,注重培养学生动手能力 、对知识的综合应用能力 、创新精神和主动探索精神[2]。

2.2 改革教学方式为"以学生为主体,以任务为载体,分组实训"。

以学生为主体,根据学生所具有的基础,进行不同的分组,选择不同的能力模块任务进行实训。在完成任务的过程中,应以学生为主体 ,以教师为主导 ,教师适度点拨指导,学生之间相互协作。学生分组相互协作完成任务时,应注意明确分工 ,促使学生间互相帮助、互相学习、互相监督。应培养学生的团队精神,减少计算机中娱乐性内容对学生的诱惑,促进他们集中精力、相互配合完成任务。当学生通过自主探究、相互讨论、分组合作等方式完成任务时，会有一种满足感、自豪感,从而更能调动、提高、激发学习新知识、新技术的欲望[3]。

学生完成实训任务后及时评价反馈。学生的学习基础、理解能力、接受能力各不相同。教师在检查学生上传的作业后,应根据其具体情况进行分析、总结,引导他们对学生的任务进行讨论，最后由教师对学生的学习情况做出反馈和评价，促进学生在交流中学会学习、学会合作。教师在评价时应对任务完成出色的学生进行表扬、鼓励,提高他们的自信心和学习热情;对学生错误较多的地方重点讲解，并将其设计为下一堂课任务的复习要点，促进学生在一种良好的心理状态下不断学习,不断进步。

2.3 改革考核方式突出基于能力模块的过程性考核方式。

本课程采用过程性评价与终结性评价相结合的评价方式,过程性评价与终结性评价各占 50%。

考核阶段	过程性评价阶段				终结性评价阶段
考核内容	平时学习表现	能力模块单元测试	综合能力测试	激励考核	期末综合考试
成绩比例	10%	20%	10%	10%	50%

表 1 考核方式

激励考核部分主要是对基础差但进步很大、有良好的团队意识,主动帮助其他同学、在计算机各种技能大赛中表现优异的同学进行鼓励。从而激发不同基础水平的学生都能主动深入的学习。

3、小结

计算机应用基础课程涉及知识面广，对实践能力要求较高,高职学生入学时基础差异较大,部分内容跟中学阶段相关课程内容重叠等，都促使高职计算机应用基础课程必须进行改革。本校的计算机应用基础课程改革实践坚持"以职业活动为导向,以能力目标为驱动,以学生为主体,以项目为载体"的思路。在教学实践中重视将"教、学、做"融为一体,强调学生实际操作能力的培养。通过以上措施的实施,有效的提高了教学质量。

参考文献：

[1] 戴士弘.职业教育课程教学改革[M].清华大学出版社,2007,(15).

[2] 戴士弘. 高职教改课程教学设计案例集 [M]. 清华大学出版社,2007,(96-119).

[3] 杨爱鑫. 信息技术教学中基于任务驱动的小组合作教学法初探[J].教育与职业,2009,(6).

(上接第 25 页)

参考文献：

[1] 王琳琳. 基于 HTML Parser 的 Web 信息提取技术 [D]. 北京邮电大学,2007.

[2] 郑长松. Web 信息智能抽取技术的研究与实现 [D]. 电子科技大学,2009.

[3] 刘艳敏 等.Web 页面主题信息抽取研究与实现. 计算机工程与应用[J],2006,(21),146-148.

[4] 于海燕 等. Web 文本内容过滤方法的研究 [J].微电子学与计算机,2006,23(9):51-54.

[5] 宋吉广 译. Python 核心编程(第二版)[M]. 北京:人民邮电出版社,2008.

[6] 王云 等.文本搜索的一种间接方法[J].四川兵工学报,2010,31(1):127-128.

[7] 周水庚 等.基于文件实例的中文信息检索[J].小型微型计算机系统,2001,22(2):14-16.