网页信息自动提取的设计与实现

栗勇兵，韩平，董启雄
（装备学院信息管理中心，北京怀柔 101416）

摘要：网页信息自动提取是一种重要的网络应用技术，用于提取各类网页的专门信息。网页信息自动提取的设计基于网页的半结构化特征，其流程是先获取 HTML 源文件，然后去掉 HTML 标记和无关信息，再进行语义匹配，提取信息到特定的数据结构，进行 CSV 格式化输出，就可以得到所需的信息。在网站企业化、网店普遍化和网络购物盛行的背景下，推广和应用网页信息自动提取技术，有着重要的经济价值和意义。

关键词：网页信息；提取；设计
中图分类号：TP311.1 **文献标识码**：A **文章编号**：1007-9599 (2012) 18-0187-02

1 引言

随着网络技术的发展和成熟，大量网站涌现，而且各种网页增生着大量的信息，各种信息集萃在诺大的网络上，形成了巨大的信息资源库。如何有目的、有选择地提取某一方面的信息，就成了一门技术，即信息提取（INFORMATION EXTRACTION）技术。信息提取是利用计算机自动从网页中收集有用的信息，充分利用信息资源的必由之路。信息提取的主要功能是从网站网页文本中提取出特定的事实信息（FACTUALINFORMATION）。比如，从经济新闻中提取出公司发布新产品的情况，包括公司名、产品名、发布时间、产品性能等；从网店中提取店铺信息、商品信息，以及服务信息等；从新闻报道中提取出某一恐怖事件的详细情况，包括时间、地点、作案者、受害者、袭击目标和使用的武器等；从病人的医疗记录中提取出症状、诊断记录、检验结果和处方等，或者直接提取网页文章中某句话或某段话的信息等。通常，被提取出来的信息以结构化的形式描述，可以直接存入数据库中，供用户查询以及进一步分析利用。

为了处理海量文本，信息提取系统通常以信息检索系统（例如文本过滤）的输出作为输入；而信息提取技术又可以用来提高信息检索系统的性能。信息提取和信息检索二者的结合能够更好地服务于用户的信息处理需求。在信息提取中，用户一般只关心有限的感兴趣的事实信息，而不关心文本意义的细微差别以及作者的写作意图等深层理解问题。因此，信息提取只能算是一种浅层的或者说简化的文本理解技术。既然如此，信息提取技术是如何从浩如烟海的网页中提取出所需的信息的呢？我们设想构建流水线式的提取流程，用 Python 语言，可以实现信息的自动提取。

2 网页信息自动提取的设计

2.1 网页的半结构化特征利于信息提取流程的建构

一般地，网站的网页具有半结构化特征，即网页包罗的信息多种多样，有着大量的文本、图片、视频等，甚至还有其它多媒体字符流的集合。人们认为这很难通过结构化查询语句来处理网页信息。加上网页中还含有大量的广告、导航等与所需主题可能不相关的信息，这些信息会干扰理解网站页面中所包含的目标意义。其次，诸多网页主要采用 HTML 格式书写而成，基本遵从 W3C 规范。且网站服务平台大多要求网站套用平台模板，页面要遵循平台提供者规定的框架规则，这些规则可以提供给信息提取者以启发性的线索。

2.2 网页信息提取流程

可.以根据网页的半结构化特征，构想信息自动提取流程。如图 1 所示。先利用网络爬虫(WEB CRAWLER)获取并保存 WEB 页面到 HTML 文件；然后通过过滤器去掉 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页信息到特定的数据结构，进行 CSV 格式化输出。这样一个提取流程采用流水线(PIPELINE)架构，以文本文件为处理媒介。

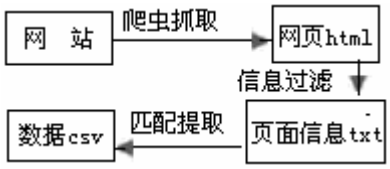

图 1 网页信息提取流程示意图

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。

3 关键技术和工具

在提取流程中，有几个重要环节需要用到专门的技术和工具。

3.1 HTML 文件的信息过滤

通用的信息提取技术有很多，我们尝试选用 HTML 标记和关键信息的正则表达式过滤方法来提取网页信息。

正则表达式是从非结构化文本中提取有用信息的利器，它通过分析网页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。一般网页用 CSS 文件将页面渲染的格式分离开来，而用 HTML 文件存放网页信息、专题信息和服务反馈评价信息等，例如一些企业网、网上商店网页等。因此，可以用正则表达式将 HTML 的主文件中的格式控制、标记等无关信息过滤掉。我们选用哪个动态语言 Python 作为实现工具。

HTML 文件过滤的关键程序代码如下：

```
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)

过程是：把代码逐行读入 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，有些网站的信用信息有专门的信用等级图标，像陶宝店网站就用红心、蓝钻、蓝冠、黄冠等来表示。因此，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。

3.2 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站网页的信息提取。索引术语技术主要假定文本库中的文件语义和用户的信息需求语义，可以用同一组索引术语来表示。而一些企业网站都有相对固定的格式，例如网络商店网页中有累计信用的信息文本片断为“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，网店名、累计信用、好评率、卖家服务态度、商品数量等重要的店铺信息都可以找到对应的索引标记，基于此，我们就可以按图索骥，利用这些索引来标识并提取有用信息。

索引术语技术的工作原理是，假设信息文本文件中的字符串为“累计信用：s_blue_3”，通过分析网店网页的信用等级信息可知，需提取的信息是“blue”和“3”。信息提取的代码如下：

①a=re.compile(u' 累计信用:\w_\w+_\w')

②if a.search(line):

③p=a.search(line).group().split(u"_")[1]

④q=a.search(line).group().split(u"_")[2]

其过程是：

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')等，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， 如企业名、企业累计信用、企业好评率、企业卖家服务态度、企业商品数量等，然后写入特定的数据结构中即可。

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。CSV 的全称是 comma-separated value，意思是“被逗号分隔的取值”。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。

4 信息提取实例

该例以网店为例，说明提取该网店信息的过程。该网店的页面局部信息如下：

描述相符：4.5 低于 4.25%;

服务态度：4.8 高于 44.71%

发货速度：4.5 低于 2.31%

店铺信息

好评：99.67% 宝贝数量：175

开店时间：2011-03-08

该网店的网页面的 HTML 文件片断，较为复杂，这里略去不附图。实际上，HTML 文件中包括了有用信息在内的大量噪声字符。经过过滤处理，就可以得到如下琐事的信息文本文件：

描述相符：4.5

服务态度：4.8

发货速度：4.6

好评率：96.67%

宝贝数量：175

开店时间：2011-03-08

再对这些信息文本文件进行匹配提取，并且按 CSV 格式进行格式化，就可以输出结构化的信息（图略）。

5 小结

网页信息自动提取的过程主要是爬虫抓取、信息过滤和匹配提取等。这一流水过程具有实用性。如果将信息提取工具作出修改，就可以用于各种企业网站平台的信息提取。这一技术，既可以用于自动搜集网站的运营信息，以监控网络购物品类；也可用以收集和跟踪竞争对手的经营情况；还可以用于对多个网站和网店的信息收集和比对商品价格、企业服务状况和信息反馈等，为客户提出购物建议等。推广和应用网站信息自动提取技术，对于网站的企业化应用，提高网络技术的实用性和经济价值，都具有重要的意义，

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.