通用论坛正文提取

一、问题的背景

在当今的大数据时代里，伴随着互联网和移动互联网的高速发展，人们产生的数据总量呈现急剧增长的趋势，当前大约每六个月互联网中产生的数据总量就会翻一番。互联网产生的海量数据中蕴含着大量的信息，已成为政府和企业的一个重要数据来源，互联网数据处理也已成为一个有重大需求的热门行业。借助网络爬虫技术，我们能够快速从互联网中获取海量的公开网页数据，对这些数据进行分析和挖掘，从中提取出有价值的信息，能帮助并指导我们进行商业决策、舆论分析、社会调查、政策制定等工作。但是，大部分网页数据是以半结构化的数据格式呈现的，我们需要的信息在页面上往往淹没在大量的广告、图标、链接等“噪音”元素中。如何从网页中有效提取所需要的信息，一直是互联网数据处理行业关注的重点问题之一。

网页通常采用超级文本标记语言（英文缩写：HTML）来编写，页面上的不同元素如作者、主题、发布日期等出现在一对特定的标记符之间。例如当我们看到如下一个论坛网页：

图 1

我们可以通过查看这个网页的源代码，查看到与之对应的信息

（1）标题信息：

```
<div class="atl-location clearfix">
        <p class="crumbs"><a href="http://bbs.tianya.cn">天涯论坛</a> &gt; <em><a
href="/list-university-1.shtml">我的大学</a> &gt; <a href="/list.jsp?item=university&sub=8"
rel="nofollow">话题互动</a></em> <a href="compose.jsp?item=university" class="addpost" rel="nofollow">[我
要发帖]</a></p>
    </div>
    <div id="post_head" class="atl-head">
        <h1 class="atl-title">
            <span class="s_title"><span style="font-weight:400;">应届毕业生，谈谈自己
工作选择的迷茫，也请路过的各位前辈不吝赐教</span></span><span id="q_button_170_25_w"><div
id="q_button_170_25"></div>
            </span>
        </h1>
```

图 2

（2）题主信息：

```
<div class="atl-info">
                <span>楼主：<a href="http://www.tianya.cn/97370897"
target="_blank" class="js-vip-check" uid="97370897" uname="[redacted]">[redacted]</a></span>
                <span>时间：2017-03-08 09:46:00 </span>
                <span>点击：24 </span>
            <span title="共10个回帖和1个评论">回复：11</span>
            </div>
    </div>
```

图 3

（3）题主发帖内容

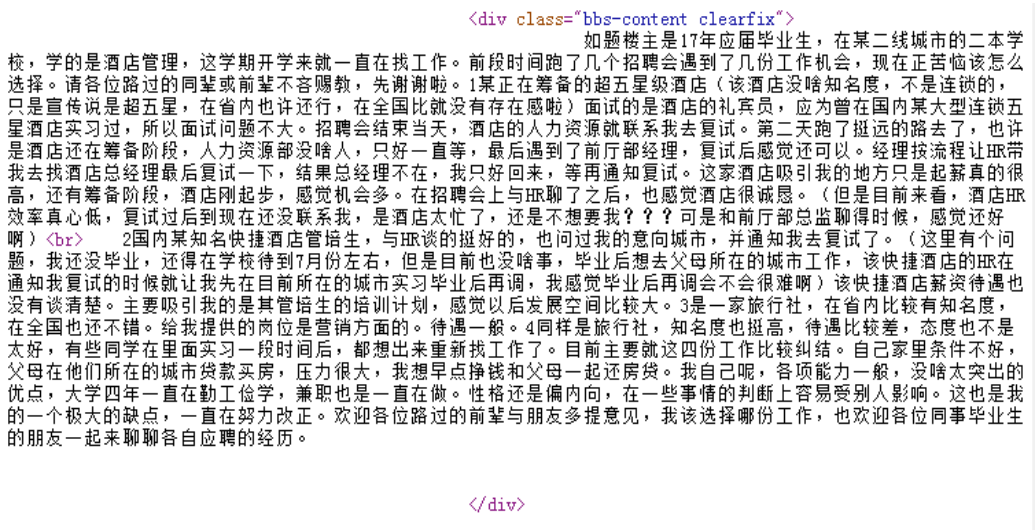

图 4

（4）回帖信息

图 5

图中的网页源代码就是超级文本标记语言（HTML），关于超级文本标记语言百度百科中是这样描述的：

超级文本标记语言是标准通用标记语言下的一个应用，也是一种规范，一种标准，它通过标记符号来标记要显示的网页中的各个部分。网页文件本身是一种文本文件，通过在文本文件中添加标记符，可以告诉浏览器如何显示其中的内容（如：文字如何处理，画面如何安排，图片如何显示等）。

维基百科中对 HTML 语言的标记、元素、属性、数据类型等也有详细的描述和样例说明。

对于给定的一个具体网页，通常的做法是，人工分析这个网页的源代码，找到特定内容对应的标签，然后通过关键字匹配（例如标签匹配）的方法就可以从网页源代码中获取到我

们所关心的数据，如下表所示：

表 1　HTML 标签与内容的对应

标题：<h1 class="atl-title">
题主：
发帖内容：
回帖信息：

但是，不同网站甚至网页所使用的网页格式、网页结构和标签体系都可能是不一样的，对于从互联网中获取的海量网页的批量处理，如果还利用传统的方法去对每个有差异的网页逐一做人工分析，是不可行的。如何从这些存在差异的网页中快速有效的提取所需信息，就成为互联网数据处理中一个急需解决的问题。

在传统的网页结构化数据提取智能分析实践中，已经有很多开源的智能提取算法来分析新闻、文章类数据，但是这些方法只适用于提取有大段文本的页面结构数据信息，如：网页的作者(author)、标题(title)、正文内容(content)、发布时间(publish_date)。对于 BBS 论坛类的网页，由于文本在网页上相对分散，提取的字段更多，传统的算法不再适用，需要重新设计通用提取算法，针对主题帖(post)和回帖(reply)进行有效地分析提取。

本赛题是针对当前互联网数据处理行业的这一实际需求而提出，旨在研究如何高效、智能地从海量论坛网页中自动地进行内容抽取，提炼出其中的有价值信息。

二、请实现以下目标

对于任意 BBS 类型的网页，获取其 HTML 文本内容，设计一个智能提取该页面的主贴、所有回帖的算法。如下面的网页截图所示，提取主贴和回帖的区域，提取出相应数据字段（只需要提取文本， 图片、视频、音乐等媒体可以直接忽略），并按规定的数据格式（Json 格式）存储。

图 6

重要说明：

1. Json 数据字段说明：
   - post ：主题帖
     - author： 用户名
     - title：标题
     - content：帖子内容
     - publish_date:帖子的发布日期，格式： yyyy-MM-dd
   - replys: 该页的回帖列表
     - 每条回帖的主要字段同 post， 若回帖无 title 字段，可为空
2. 算法要求：
   - 算法必须具有通用性，必须支持互联网的任意类型 BBS 网站，不得只针对附件所给的样例网站、或特定类型的开源论坛（例如 discuz、phpwind）

三、数据样例

1. 样例输入数据格式：

（每行一条论坛的内容页的 url）

http://bbs.tianya.cn/post-stocks-1841155-1.shtml （包含主贴的 url)

http://bbs.tianya.cn/post-stocks-1841155-3.shtml （不包含主贴的 url）

2. 样例输出数据格式（必须按如下格式提交结果）：

每行数据有 {原始 url}\t {提取结果的 json 字符串}， 表示某一个 html 页面(url)提取出来的数据，示例数据格式：http://x.heshuicun.com/forum.php?mod=viewthread&tid=80

{"post":{"content":"本帖最后由 临风有点冷 于 2012-7-27 08:26 编辑 从这条新闻中你得到了什么教训 在网上看个新闻，大概内容是：老公买了一只藏獒幼仔，没时间养，一直是老婆在养;一次老公老婆吵架,老公把老婆打了,结果藏獒冲出来果断把老公手咬断了！看完新闻我问老公:“从这条新闻中你得到了什么教训？” 本想听他说不能打老婆，没想到这货居然说:“ 游客，如果您要查看本帖隐藏内容请回复” ","title":"从这条新闻中你得到了什么教训？ ","publish_date":"20120727"},"replys":[{"content":"的分数高如果认购二哥让他退给我","title":"从这条新闻中你得到了什么教训？ ","publish_date":"20160904"},{"content":"1231231231321321","title":"从这条新闻中你得到了什么教训？ ","publish_date":"20161115"},{"content":"","title":"从这条新闻中你得到了什么教训？ ","publish_date":"20161208"}]}

附件 1：bbs_urls.txt

附件 2：result_sample.txt