第五届泰迪杯
数据挖掘挑战赛

优秀作品

作品名称：通用论坛正文提取

荣获奖项：特等奖

作品单位：华南师范大学

作品成员：徐文扬 孙孟涛 陈天琪

指导老师：刘磊

通用论坛正文提取

摘　要：

在当今的大数据时代里，当前每六个月互联网中产生的数据总量就会翻一番。借助网络爬虫技术提取数据资料至关重要。并且网页论坛的结构各种各样，能够对于多样论坛有一个通用提取算法，则是可以快速获取的数据的有利条件之一。

本文完成了对主题帖标题，作者，发帖时间和主题帖正文的提取，以及回帖楼层，作者，回帖时间和回帖正文的提取。经过一个月的程序调试，最终得到了得到了较为完整的爬虫内容，特别是在处理噪声内容方面，达到了较好的结果。

本文第一章简单介绍本题背景并说明文章应解决的问题；第二章说明本次网页爬虫流程；第三章主要针对本爬虫算法中的网络技术进行简要分析；在第四章中详细地介绍了本次爬虫的具体算法。在 4.1 节中主要解决的是主题帖内容的获取，包括主题帖标题、作者、发帖时间以及主题帖正文，其中设计特征词库，设计噪声词库，在提取时间时应用了时间正则和特征提取算法，在提取作者时应用了标签特征法、URL 相似度验证法和噪声过滤法，之后判断网页结构，提取正文等。在 4.2 节中主要解决的是回帖内容的获取，包括回帖作者、回帖时间以及回帖正文，其中设计了定位楼层的算法，根据定位楼层的算法，不仅可以得知每个回帖具体的楼层信息，也可以智能获取每个回帖标签的共同特征。应用这些特征，可以对主题帖爬取内容进行修正，同时为爬取相同的论坛的文本提取奠定了基础。第五章的内容主要是针对真实网站进行数据提取的展示，将对三个网站进行文本提取。之后第六章的内容是对本次算法的补充以及实验数据分析。

本文所涉及爬虫算法较为完整地得到了论坛数据且去掉了足够多的噪声数据，其中时间同步回溯算法均是拿到题目之后一次次尝试所得到的，并且对于一些小细节的处理足够精致，最后的源程序也说明了本次算法得到了较为成功的提取算法。

关键词：Beautifulsoup　正则表达式　网页结构　作者 URL 特征法　噪声过滤

目 录

1. 挖掘目标

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

能够实现：

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。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。

2. 全文脉络图

本次网页爬虫主要流程如下：

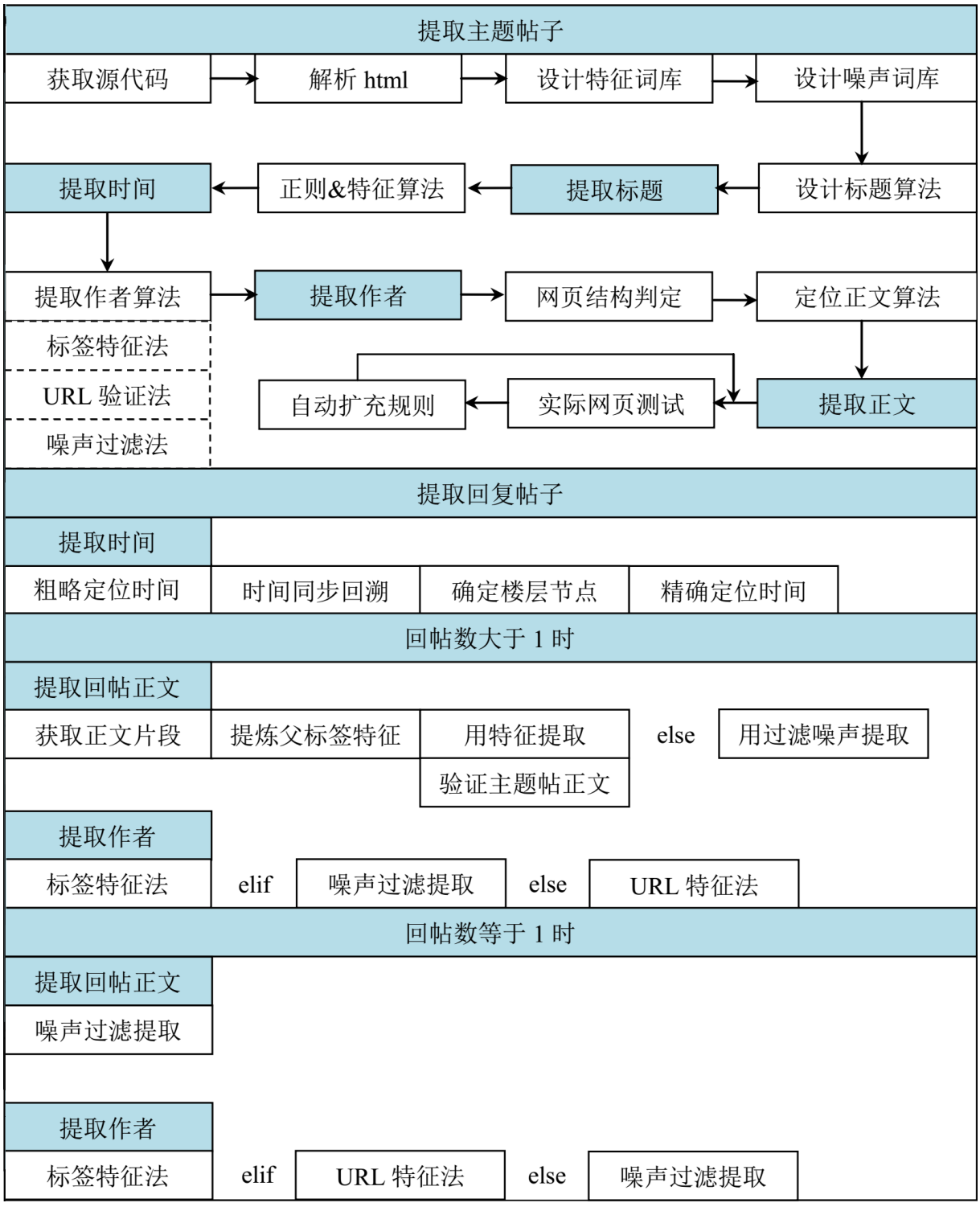


图 1 全文脉络图

3. 爬虫技术简介

3.1. 爬虫简介

网络爬虫是一个自动提取网页的程序，它为搜索引擎从网页上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

网络爬虫分为以下几种类型通用网络爬虫（General Purpose Web Crawler）、聚焦网络爬虫（Focused Web Crawler）、增量式网络爬虫（Incremental Web Crawler）、深层网络爬虫（Deep Web Crawler）。本文所设计的算法即是通用网络爬虫，针对所有论坛，可以爬虫到帖子的作者，发帖时间，发帖内容，回帖作者，回帖时间以及回帖内容。

3.2. 正则表达式介绍

正则表达式的“鼻祖”或许可一直追溯到科学家对人类神经系统工作原理的早期研究。美国新泽西州的 Warren McCulloch 和出生在美国底特律的 Walter Pitts 这两位神经生理方面的科学家，研究出了一种用数学方式来描述神经网络的新方法，他们创造性地将神经系统中的神经元描述成了小而简单的自动控制元，从而作出了一项伟大的工作革新。

经过一系列发展，正则表达式在各种计算机语言或各种应用领域得到了广大的应用和发展，演变成为计算机技术森林中的必不可少的应用工具。

正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。

给定一个正则表达式和另一个字符串，我们可以达到如下的目的：对于给定的字符

串是否符合正则表达式的过滤逻辑（称作“匹配”）；以及可以通过正则表达式，从字符串中获取我们想要的特定部分。

正则表达式的特点是：灵活性、逻辑性和功能性非常的强；可以迅速地用极简单的方式达到字符串的复杂控制。对于刚接触的人来说，比较晦涩难懂。

并且由于正则表达式主要应用对象是文本，因此它在各种文本编辑器场合都有应用，小到著名编辑器 EditPlus，大到 Microsoft Word、Visual Studio 等大型编辑器，都可以使用正则表达式来处理文本内容。

4. 具体步骤

4.1. 解题思路

题目要求给出各类论坛的通用爬虫，由于论坛建设者的英文标签起名多样化，细节的结构复杂，传统爬虫根据标签的爬取方法不再适用。为设计通用爬虫，首先应该找到对象的共性。对于论坛网站，常见规律分析如下：

1.呈现给读者的文字顺序通常是有规律可循的，如：标题，会出现在时间，作者，正文之前，且通常不会太远；

2.中文的一些命名习惯常常是通用的，如：注册，登录，积分，举报，回复，等等噪声词汇；

3.时间通常会精确到分钟甚至到秒，且时间的格式较为统一；

4.大多数带有超链接的文字通常不会是正文，可以进一步的去除噪声；

5.统一论坛的不同作者空间，其 url 通常是有很多重叠部分。

6.回帖楼层的布局通常是相似的。

基于以上规律，本文提供的算法将通过楼层的相似性分析定位楼层，以过滤噪声词汇的方式找到部分有效信息，对比其在 dom 树中的位置进而精确定位有效信息。对于无回复贴的情况，将需要结合标题，作者，时间，正文的位置规律，结合 dom 树的结构，更好的过滤掉噪声信息。

算法的核心是过滤噪声信息，从而通过部分有效信息精确定位。为次，我们将论坛的帖子大致分为两种类型，了解其噪声信息分布，针对性的去燥。一种是作者及帖子的正文分为左右两边即由左至右，一种是作者及帖子的内容为由上至下的内容。针对这两

种论坛的结构设计了相应的算法，可以自动区分论坛结构进行相应的爬取。

图 2 论坛结构为由上至下

图 3 论坛结构为由左至右

对于回复较多的帖子，本文通过相似性分析，寻找结构相似的节点，精确定位楼层以及所需要的信息，并给予一些强验证规则，若验证失败、未找到相似性或没有回复时，则通过噪声过滤法获得所需内容。

本文对论坛帖子作者的发帖时间进行提取，之后提取作者，正文，以及 回帖内容。看似简简单单地一步步流程，但是在考虑每一步时均会出现噪声内容的干扰，在提取有效信息时，我们建立了标志词库和噪声词库并根据论坛语言习惯，设计了很多删除以及截取规则等等。下面的开始，详细地向各位展示我们的程序思路。

4.2. 提取主题帖

4.2.1. 提取标题

大家进去论坛，首先是被论坛的帖子所吸引，所以爬虫帖子的信息也是从论坛帖子的标题开始的。一般论坛标题的位置是在源代码<head>标签下的<title>，但是不全是标题文字，还会包含其他论坛信息比如一个社区帖子标题 分类以及论坛名字，但是这几部分内容大多是以相同的分隔符区分开，多是“-”、“|”、“_”。一般首先出现的是标题，所以第一个分隔符前的是标题，这是寻找标题的第一种方法。若此方法失效，一般能完成 90%以上的标题提取。则算法自动运行第二段程序。

图 4 算法一标题源代码位置展示

第二段算法主要是，建立标题特征词库，如：{“标题”， “主题”， “题目”}寻找<body>标签下的<h1>,<h2>,<h3><title>中非空文本的标签，用正则表达式判断标签<h1>,<h2>,<h3><title>的文本是否包含标题特征词库中的词，如果包含，取出该特征词所在文本之后的文本，如果寻找特征词寻找失败，则返回第一个匹配到的长文本串，因为大多论坛标题均在开始位置出现。

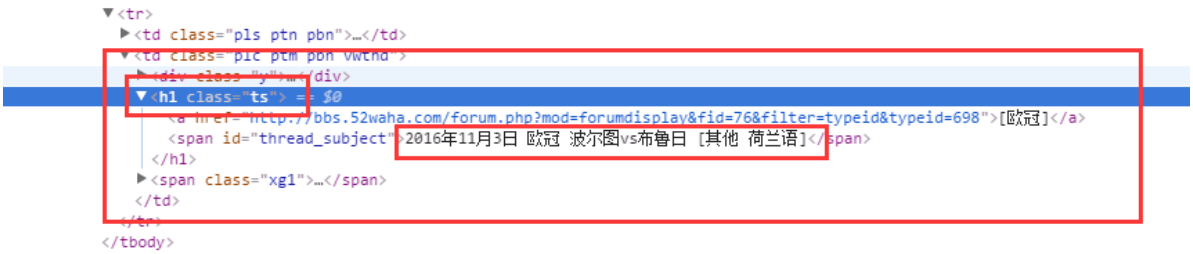


图 5 算法二标题源代码位置展示

4.2.2. 提取时间

论坛的新闻热度是靠发帖时间来判断的，通常越流行的事物发帖时间越靠近现在的时间点，而一件事情过去很久，贴子的热度仍旧不散说明这件事情很有研究价值。所以发帖时间的提取是至关重要的。首先删除源代码的空白字符。大多数论坛的时间表示均遵循一定规律：XXXX-XX-XX XX-XX-XX 或 X 天（小时，分钟）前，

所以可以使用正则表达式：

[0-9]{2,4}[-年/][0-9]{1,2}[-月/][0-9]{1,2}日*[0-9]{1,2}:[0-9]{1,2}:*[0-9]{0,2}

[今昨前 1-9] *[天小分][时钟]*前* *[0-9]{0,2}:*[0-9]{0,2}:*[0-9]{0,2}

匹配时间，但是所有出现的时间并不一定是发帖时间，还可能出现用户登录时间，注册时间，更新时间等噪声。所以需要建立时间特征噪声词典，出现在词典中的字符串均属于时间噪声数据，通过词典过滤匹配到的第一个时间即为帖子发帖时间。

4.2.3. 提取作者

作者的提取主要依据三种方法：标签特征法、作者 url 特征法、噪声过滤法。

标签特征法：发帖人的英文通常会用 author 和 username。因此我们优先寻找标签名或属性包含正则表达式^auth|^us*e*r*_*name 的标签，提取其文本内容，这样提取出的作者通常是准确的，我们通过超对比作者超链接的方式，进一步验证提取到的准确性。

图 6 回复帖作者超链接网址相似展示图

url 特征法：通常情况下，论坛设计者为了游客更好的了解作者，会将作者名设置为一个指向其论坛空间的超链接，而对于不同的作者，他们的 url 通常是有共性的，我们遍历源代码中的其它 url，计算其和作者 url 的最长公共字符串的长度。通常情况下，url 中作为作者标识的内容不会超过 12 位，因此若公共长度与作者 url 长度相差超过 12 位的，我们通过标签特征名得到的放弃结果。另一方面，为了防止回帖人与发帖人标签特征不一样，若公共长度与作者 url 长度相差小于 12 位的，根据公共部分作为作者 url 的特征，找到的第一个满足的 url，用其标签的文本内容作为主题帖作者修正前面得到的结果。

若上面的方法没有得到作者，我们将采取噪声信息过滤的方法。具体步骤如下：

噪声信息过滤法：将源代码<body>内的不同标签的文本，按照\n 顺序分隔，并且去掉前后空格和空行进行提取，构成文本列表。

在源代码中，标题出现的位置不止一处，但通常会出现在作者和时间之前很近的位置。因此，我们优先取出现在时间之前最近的标题，寻找其在列表中的位置作为起点，逐一向下检索每行文本是否为噪声信息，直到抓取到第一条有效信息便作为作者。过滤规则如下：

1、 包含标题和时间的文本不会是作者

2、 对于含有数字的行，判断其是否含着，楼，次，个，查看，回复等作者噪声词

3、 对于含着“：:”的行，若“：:”前有楼主、作者、发帖人等作者标志词，选取“：:”后面的内容提取为作者

4、 其余情况去除噪声词汇和标点符号后，剩余 1 个字以上的，则把该行作为作者。

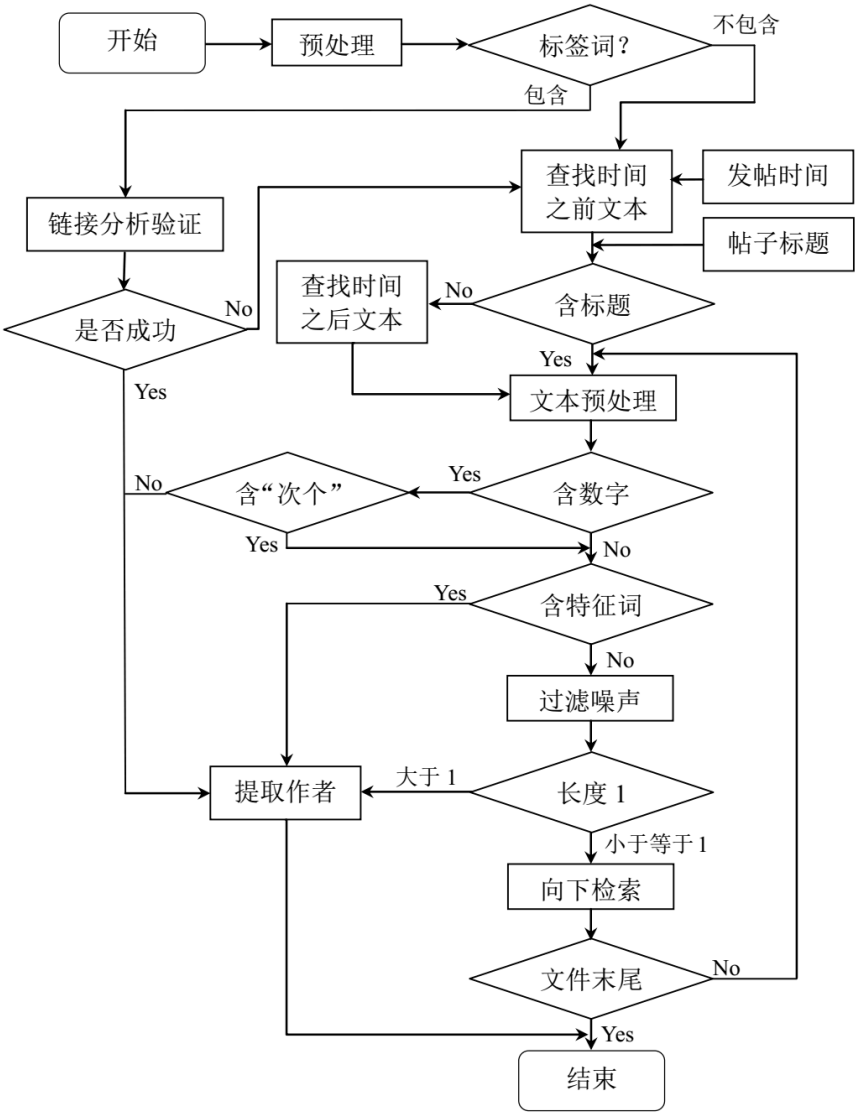


图 7 主题帖提取作者流程图

4.2.4. 判断网页结构

主题帖缺乏对照，因此噪声信息定位对正文提取的精确性至关重要。由于正文通常不会是超链接，故在过滤超链接词汇的基础上，非链接的噪声信息成为过滤的关键。而不同类型的论坛，这类噪声信息的分布各有特色。上下结构的论坛，时间和作者通常在同一区域，距离较近，且在该区域会包含少量非链接的噪声词。左右结构的论坛，在时间和作者中间夹杂着作者的论坛信息。基于这点， 我们需要先判定出论坛类型，其主

要思想是根据时间与作者的最小间隔来判断。具体算法如下：

1、 将源代码<body>内的不同标签的文本，按照\n 顺序分隔，并且去掉前后空格和空行进行提取，构成文本列表

2、 找到发帖时间第一次出现的位置

3、 根据之前提取的主题帖作者词，找到此作者词处在在本列表的所有位置

4、 定位出离发帖时间最近的主题帖作者的位置

5、 若两者距离相差不超过 4 个词，则判断此论坛为由上至下论坛格式，否则为由左至右的格式。

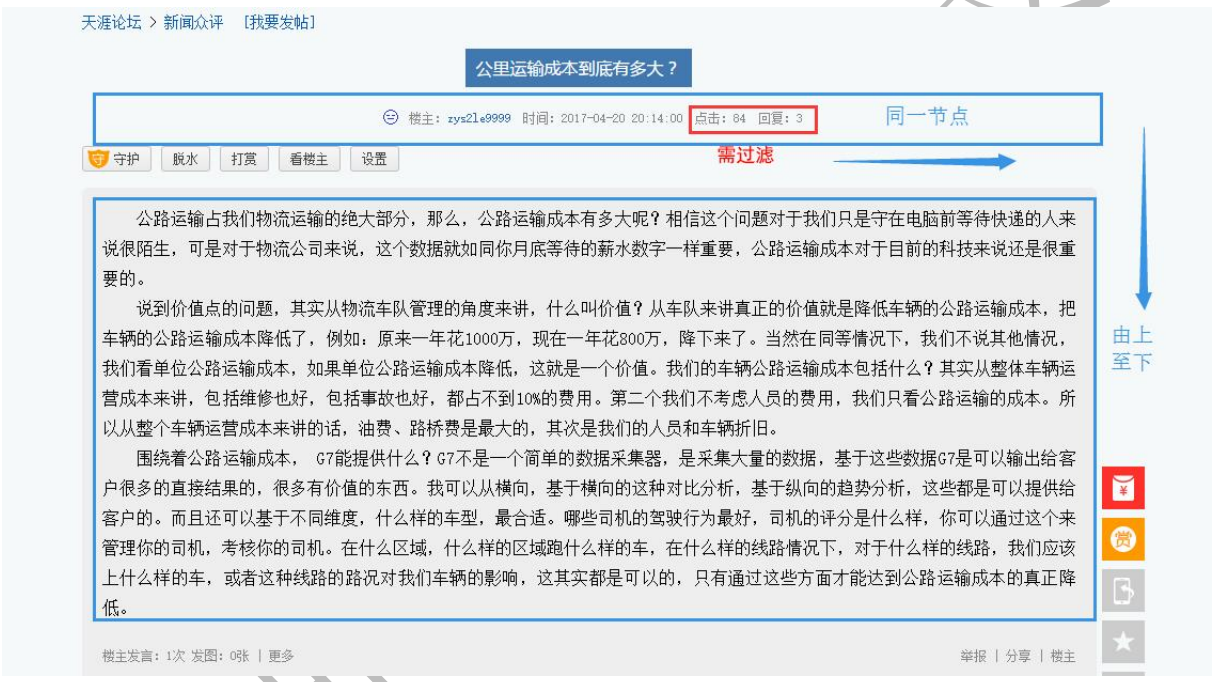


图 8 论坛结构为由上至下

图 9 论坛结构为由左至右

4.2.5. 提取正文

论坛结构的两种类型，提取正文片断方法，略有不同，详述如下：

1 论坛结构为由上至下（作者与时间相差不超过 4 个词）：

将源代码中第一次出现时间之后的文本进行提取，将不同标签的文本，按照\n 顺序分隔，并且去掉前后空格和空行，提取所有超链接的内容作为噪声信息。此外，这类论坛作者和时间通常会放在一个 dom 树的同一个节点下（加图加说明），我们从作者所在节点，逐层遍历其父节点，若在两层父节点内发现时间，则把该父节点所包含的信息都加入噪声词，否则放弃寻找。

完成上述预处理后，开始根据设定的规则逐行排查，直到出现第一行正文便停止搜索，用该行作为正文的模糊查找范围，用来进行精

图 10 噪音词汇云图

确定位文本。

规则如下：

1.如果文本含@符号，则作为正文片断。

2.若此行包含“：”，则看“：”前面的字数，若字数小于等于 4，则查找下一行。

3.若此行恒等于超链接词语，则查找下一行。

4.若包含主题帖作者，标题，发帖时间的信息，则查找下一行。

5.其他情况，若去掉“噪音”词汇、数字以及标点符号，剩余文本大于三个字，则保留作为正文片断。

2 论坛结构为由左至右（作者与时间相差多余 4 个词）

由于左右分隔的论坛，左栏会包含大量作者信息，例如积分，注册日期，登录日期，头衔等大量的“噪音信息”，因此需要定位正文所在的右栏的父标签（加图加说明），定位方法如下：

经过多方考量。左栏通常为作者信息，因此作者名必定存在，右栏通常为发帖时间以及正文信息，因此以发帖时间为起点，依次查找其父节点是否包含作者名，重复此过程，直到查找到作者名，便可定位出右栏所对应的的根节点。查找正文方法与前者算法相同。

将两种论坛的正文片断提取结束，则可以提取正文全部，提取方法如下：

对于无楼层对比的帖子，

在源代码中寻找包含正文片断的标签，寻找其父节点直到包含发帖时间为止，提取其文本内容，截取上一步得到正文片断以后的内容，并删除单独出现的噪声信息。

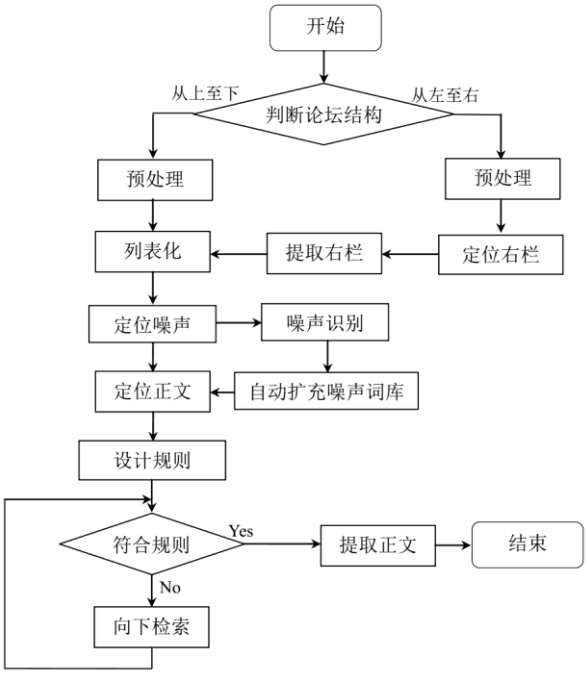


图 11 主题帖提取正文

4.3. 提取回帖

4.3.1. 定位楼层

本文的核心是噪声信息的过滤与相似结构的对比，定位楼层对于缩小回复贴内容寻找范围至关重要。此外，通过对比定位出的楼层，可以更有效的过滤掉楼层间重复性较高的噪声信息，甚至可以找出楼层的结构特征，这对准确率会有极大地提升。

由于论坛格式的不同，部分论坛主题帖和回复帖的格式一样，为了能包含更多的情况，假设主题帖和回复帖的格式不同。具体算法如下：

1. 提取出所有时间

2. 任取两个时间，同时逐层向上寻找其父节点，直到拥有相同父节点为止。

3. 若未发现相同父节点，则取其他一组时间点，重复步骤 2

4. 当全部组合均未找到相同父节点，而时间超过 1 个，则判定楼层层数为 2，回帖数目为 1。

5. 若找到相同父节点，则判定回复帖数大于 1，楼层数大于 2。此时比较回复帖根节点标签的共性，寻找与其结构相似的节点，即为全部回复帖的根节点。

6. 将找出的回复贴根节点按照出现顺序确定楼层顺序和层数。

算法的亮点是，当回复帖数大于 2 时，可以找到回复帖标签特征，并且可以直接应用到该论坛其他回复帖数目不足 2 的帖子，增加信息爬取的准确性。

4.3.2. 楼层数为 2（回复数为 1）

4.3.2.1. 提取时间

将主题帖正文后的源代码利用主题帖提取时间的方式，将第一个提取出的时间作为回复帖 1 楼的时间。

4.3.2.2. 提取作者

首先通过作者的标签特征法去寻找作者，若未发现采取 url 特征提取法，首先提取出源代码中所有的超链接，将其与主题作者的超链接做比较，选取相似度最高的链接（若有多个相似度相同的链接，则取第一个链接）将其视为回复帖 1 楼的作者。若 url 特征提取法依然无法得出结果，则从正文后用噪声过滤法。

4.3.2.3. 提取正文

提取正文方法与主题帖提取正文方法类似，唯一区别在于从回复帖 1 楼发帖时间开始遍历。

图 12 论坛帖子楼层数为 2 展示

4.3.3. 楼层数为 3 以上

由于定位楼层时已经精确定位出各回复贴根节点，故只需分别对各个节点提取有效

信息，相对主题帖来说，回复贴的噪声信息少，精确度相对较高，楼层较多时很容易通过算法找到其标签共性特征，最后将用这些共同特征去反爬主题帖正文，修正主题帖正文提取的结果。

4.3.3.1. 提取时间

利用主题帖提取时间的方式，分别在各回复帖根节点中将第一个提取出的时间作为回复帖的时间。

4.3.3.2. 提取作者

首先通过作者的标签特征法在各回复帖根节点去寻找作者，若未发现则堆各楼层根节点采取噪声过滤法获取作者，由于 url 特征提取法取决于主题帖作者提取的正确率，故最后依然无法获取作者时使用。

4.3.3.3. 提取正文

提取回复帖正文方法如下：首先针对各回复帖根节点，应用主题帖正文提取的方法，各自提取出正文片断。由于增加了楼层对比，我们将楼层间重复率较高的词语加入筛选规则，从而增加了正文片断提取的准确性。

对回复帖根节点搜索包含正文片段的标签，对比标签属性的公共部分作为正文标签特征，应用其提取全部正文内容并尝试用其提取主题帖正文。若特征提取不成功，则在各回复帖根节点中寻找对应的正文片断标签，逐层寻找各自父节点直到包含发帖时间为止，提取出文本内容，截取正文片断以后的内容，并删除单独出现的噪声信息。

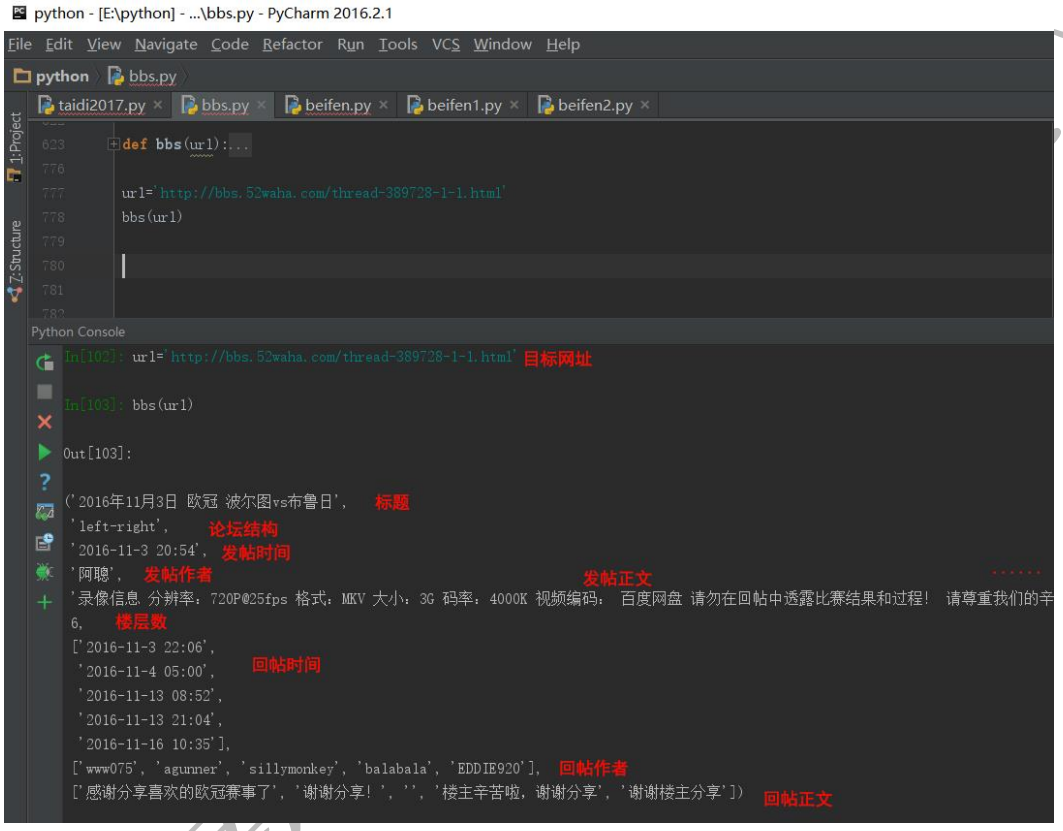
图 13 论坛帖子楼层数为 3 以上展示

5. 效果展示

为了让大家更为清晰地看到本文算法的爬虫结果，特地在十大论坛中找到三个论坛进行效果展示，三个网站分别是哇哈体育论坛、新浪论坛以及天涯论坛：

5.1. 哇哈体育论坛爬虫结果

目标网站：http://bbs.52waha.com/thread-389728-1-1.html
爬虫结果展示：

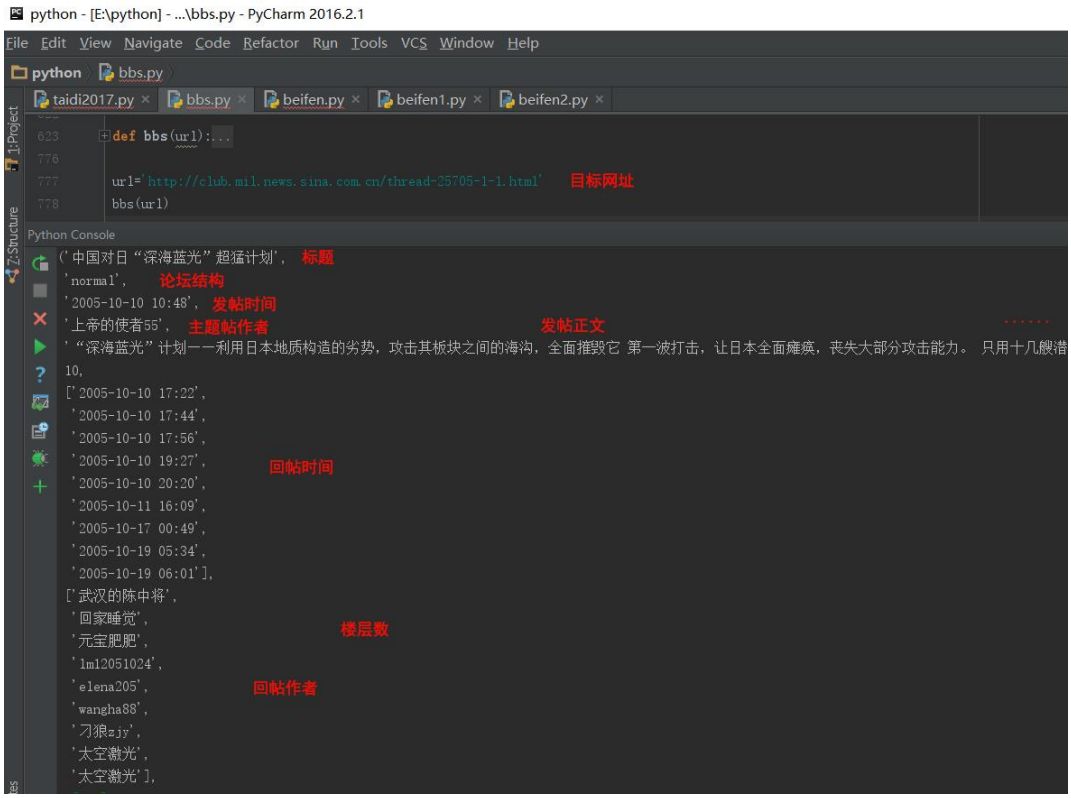


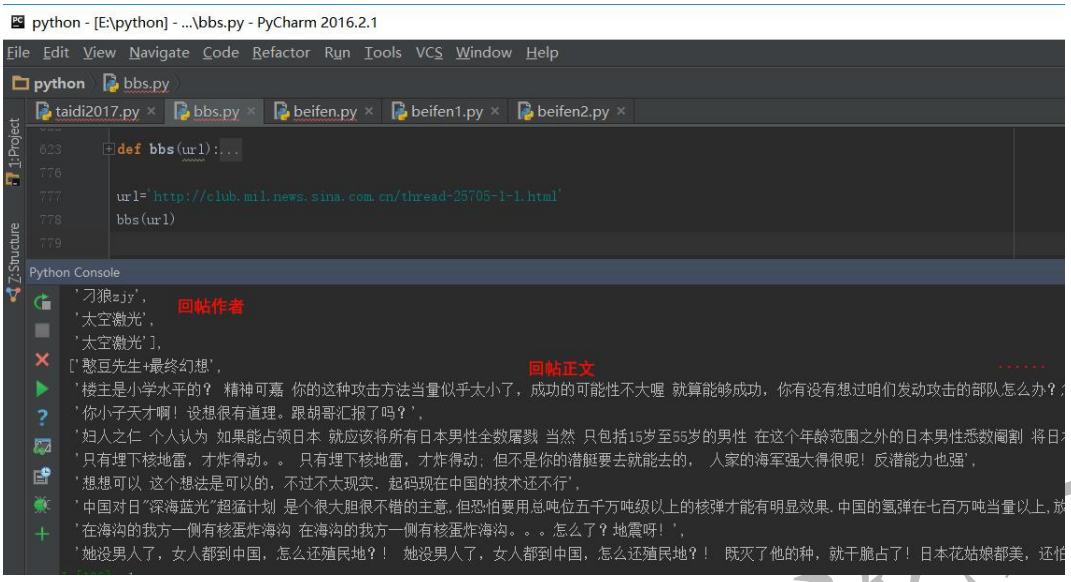
可以从上面两组图的对比得知主题帖标题、论坛结构、发帖时间、主题帖作者、楼层数、回帖作者、回帖时间、回帖正文提取的正确率极高，且噪声词汇去除较准确，且给出了正文标签的共同特征（如下图所示）

5.2. 新浪论坛爬虫结果

目标网站：http://club.mil.news.sina.com.cn/thread-25705-1-1.html
爬虫结果展示：


武汉的陈中将
回帖作者
发帖 40 精华：0 注册时间：2005-6-3 发短消息
回帖时间
楼层
2楼
憨豆先生+最终幻想
回帖正文
回复
引用
TOP
回帖作者
回家睡觉
回帖时间
发帖 0 精华：0 注册时间：2005-7-5 发短消息
楼层
3楼
楼主是小学水平的？
精神可嘉
你的这种攻击方法当量似乎太小了，成功的可能性不大喔
就算能够成功，你有没有想过咱们发动攻击的部队怎么办？怎么撤离？？？
应该是预设核弹才对
回帖正文
回复
引用
TOP
回帖作者
元宝肥肥
回帖时间
发帖 15 精华：0 注册时间：2005-6-26 发短消息
楼层
4楼
你小子天才啊！设想很有道理。跟胡哥汇报了吗？
回帖正文
回复
引用

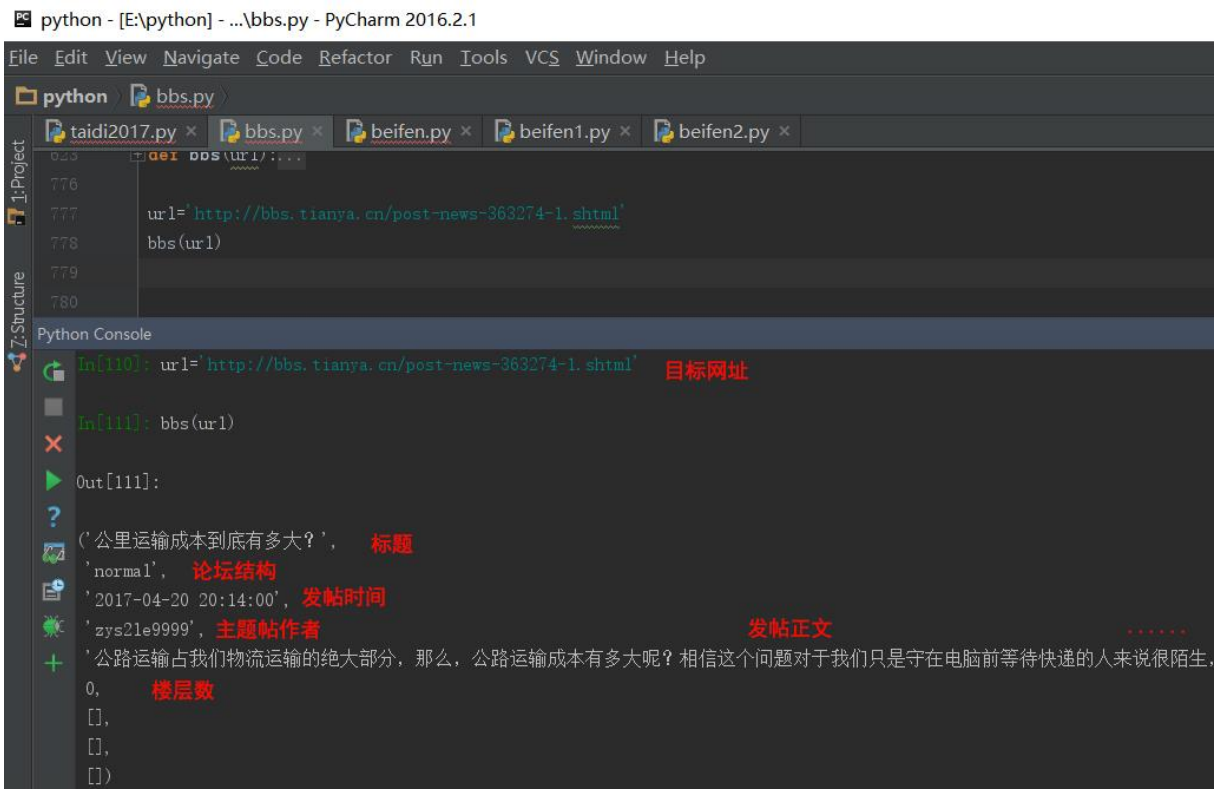

python - [E:\python] - ...\bbs.py - PyCharm 2016.2.1
File Edit View Navigate Code Refactor Run Tools VCS Window Help
python bbs.py
taidi2017.py bbs.py beifen.py beifen1.py beifen2.py
def bbs(url):...
url='http://club.mil.news.sina.com.cn/thread-25705-1-1.html'
目标网址
bbs(url)
Python Console
('中国对日“深海蓝光”超猛计划', 标题
'normal', 论坛结构
'2005-10-10 10:48', 发帖时间
'上帝的使者55', 主题帖作者
发帖正文
'“深海蓝光”计划——利用日本地质构造的劣势，攻击其板块之间的海沟，全面摧毁它 第一波打击，让日本全面瘫痪，丧失大部分攻击能力。 只用十几艘潜
10,
['2005-10-10 17:22',
'2005-10-10 17:44',
'2005-10-10 17:56',
'2005-10-10 19:27',
回帖时间
'2005-10-10 20:20',
'2005-10-11 16:09',
'2005-10-17 00:49',
'2005-10-19 05:34',
'2005-10-19 06:01'],
['武汉的陈中将',
'回家睡觉',
楼层数
'元宝肥肥',
'lm12051024',
'elena205',
回帖作者
'wangha88',
'刁狼zjy',
'太空激光',
'太空激光'],



可以从上面两组图的对比得知主题帖标题、论坛结构、发帖时间、主题帖作者、楼层数、回帖作者、回帖时间、回帖正文提取的正确率极高，噪声词汇去除较准确，且给出了正文标签的共同特征，如下图所示。

5.3. 天涯论坛爬虫结果

目标网站：http://bbs.tianya.cn/post-news-363274-1.shtml
爬虫结果展示：



可以从上面两组图的对比得知主题帖标题、论坛结构、发帖时间、主题帖作者、楼层数的正确率极高，且噪声词汇去除较准确，并且在无回帖的情况下提取无失误。

6. 参考文献

[1] Jiawei Han, Micheline Kamber, Jian Pei 著 范明 孟小峰译 数据挖掘—概念与技术 北京：机械工业出版社 2012 年

[2] Ryan Mitchchell 著 陶俊杰，陈小莉译 Python 网络数据采集 北京：人民邮电出版社 2016 年

[3] Richard Lawson 著 李斌译 用 Python 写网络爬虫 北京：人民邮电出版社 2016 年