基于Python爬虫技术的网页数据抓取与分析研究

熊畅
(三峡财务有限责任公司,北京 100038)

摘要:基于Python爬虫技术简单易用的特点,利用python语言编写爬虫程序对国家广播电视总局电视剧电子政务平台的电视剧备案数据进行了爬取。并对爬取的电视剧备案数据进行了统计分析,得出相关结论。

关键词:Python;爬虫;数据分析
中图分类号:TP311.11 **文献标识码**:A **文章编号**:1007-9416(2017)09-0035-02

1 爬虫技术简介

网络爬虫,是一种通过既定规则,自动地抓取网页信息的计算机程序。爬虫的目地在于将目标网页数据下载至本地,以便进行后续的数据分析。爬虫技术的兴起源于海量网络数据的可用性,通过爬虫技术,我们能够较为容易的获取网络数据,并通过对数据的分析,得出有价值的结论。

Python语言简单易用,现成的爬虫框架和工具包降低了使用门槛,具体使用时配合正则表达式的运用,使得数据抓取工作变得生动有趣。

2 案例分析

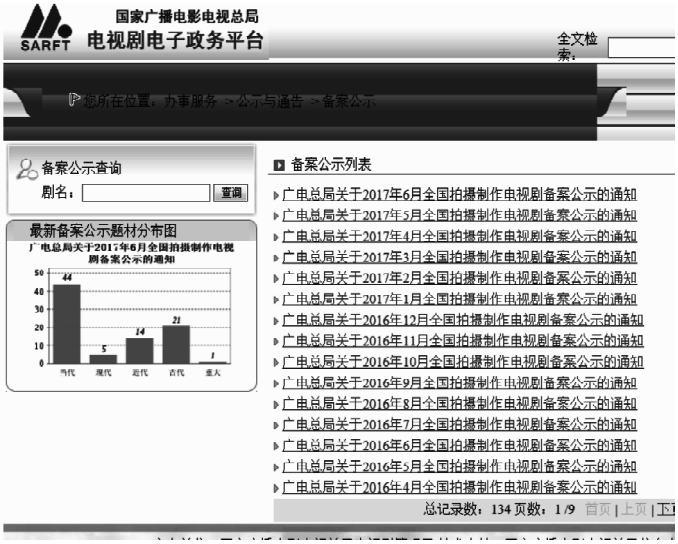

图1 目标网页信息

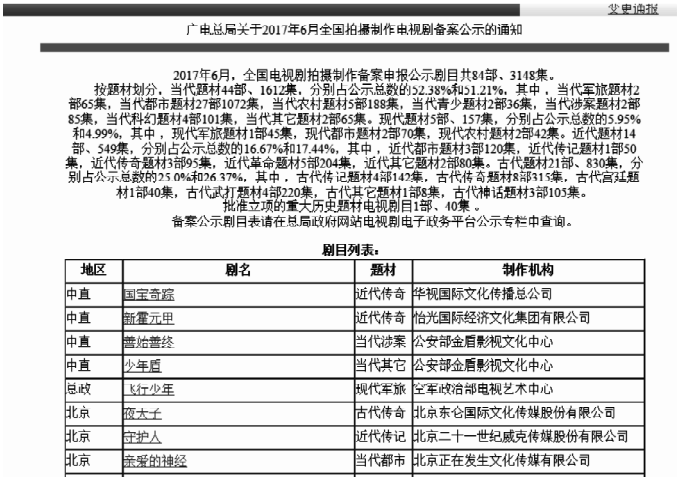

图2 表格数据信息

```
opera_data.head()
```

	制作机构	剧名	地区	年份	月份	题材
0	中国电影股份有限公司	胡同世家	中直	2017	7	现代都市
1	中国民族音像出版社	都是一家人	中直	2017	7	当代其它
2	中国电视剧制作中心有限责任公司	苍龙聚紫禁出	中央台	2017	7	古代其它
3	八一电影制片厂	大国飞天	总政	2017	7	当代其它
4	空军政治部电视艺术中心	雷霆突击	总政	2017	7	当代军旅

图3 爬取结果(表头)

```
opera_data.tail()
```

	制作机构	剧名	地区	年份	月份	题材
65	西安中大影视文化有限公司	绿色大沉浮	陕西	2009	3	当代农村
66	西安合信时代影视文化传播有限公司	下南洋	陕西	2009	3	近代传奇
67	天山电影制片厂	新疆古丽	新疆	2009	3	当代其它
68	北京华亿联盟文化传媒投资有限公司	爱在苍茫大地间	北京	2009	3	现代其它
69	西安兄弟时代影视文化传播有限公司	丐侠传奇	陕西	2009	3	近代传奇

图4 爬取结果(表尾)

2.1 网页说明

目标数据是历年来的全国电视剧拍摄备案数据。数据源于国家广播电视总局电视剧电子政务平台的公开信息,如图1所示,具体网址URL:“http://dsj.sarft.gov.cn/tims/site/views/applications.shanty?appName=note”。

我们需要爬取历年来每个月的备案公示信息列表数据,如图2,并进行汇总和分析。

2.2 爬虫程序设计并实现

首先,我们用BeautifulSoup解析器来解析URL的文本信息,分析网页HTML文本和页面规则后,制定以下步骤来抓取目标数据。

①抓取首页码和尾页码后,循环抓取列表页信息;

②通过”th”标签来提取表头信息;

收稿日期:2017-09-05

作者简介:熊畅(1983—),男,汉族,湖北黄冈人,硕士研究生,经济师,研究方向:数理金融。

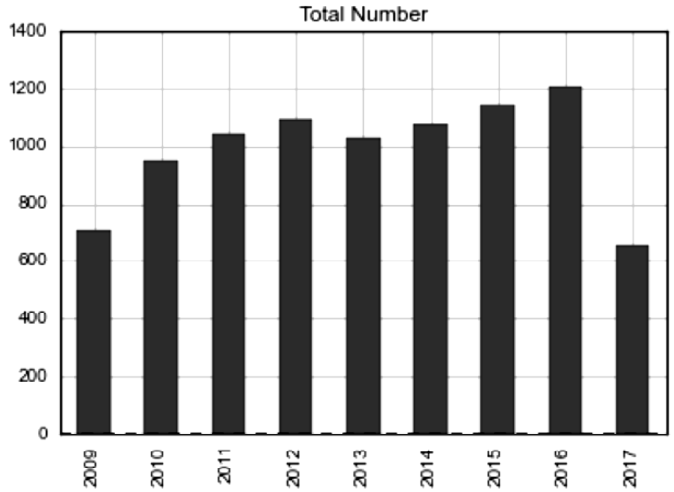

图5 年度数据

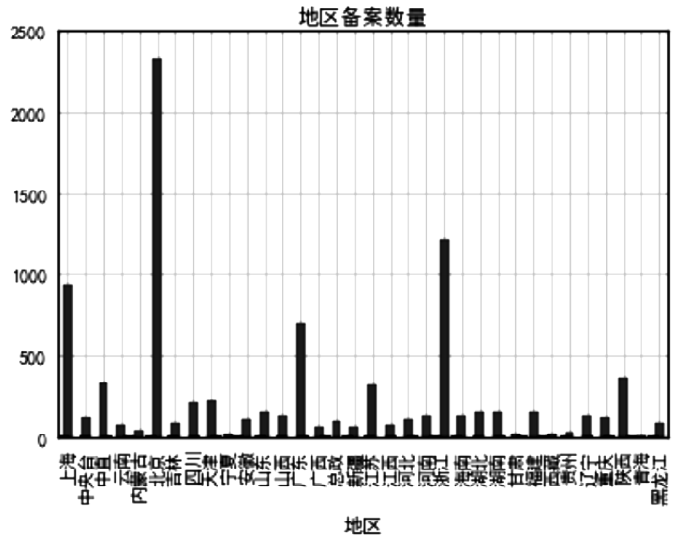

图6 地区分布

③循环提取行信息；

④将每一行的信息加入年份和月份属性，将所提取的信息组合成DataFrame格式。

用Python编程实现上述步骤，最终的结果是抓取并形成了一个8884行、6列的二维表，包含了从2009年3月份至2017年7月份的电视剧拍摄备案数据，如图3和图4所示。

2.3 数据分析

根据上述数据，我们可以运用Python的统计方法，对数据进行简单的统计和分析。

2.3.1 统计每年的拍摄数量

用groupby方法统计每年的电视剧数量并作条形图。从数据上

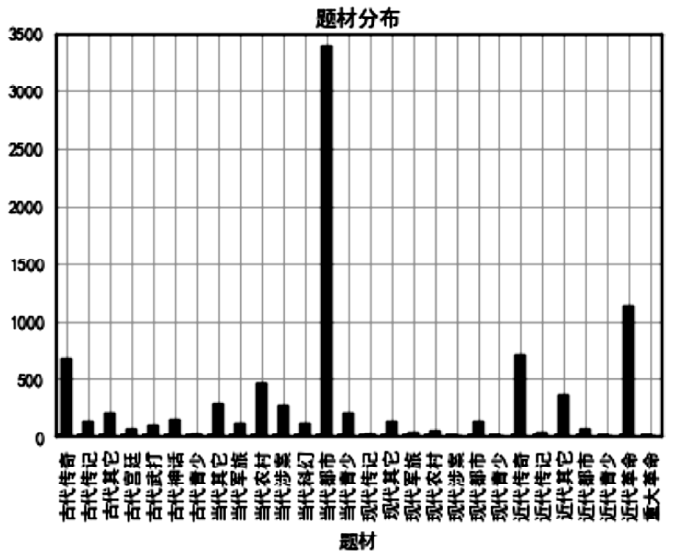

图7 题材分布

可以看出，2010年至2016年，我国电视剧备案数量整体上看呈上升趋势，如图5，从2010年的962部上升至2016年的1217部。

2.3.2 统计各地区的备案情况

同样的，用groupby方法统计各地区的备案数量。如图6所示，2009年3月份至2017年7月份，北京、浙江和上海这三个地区的电视备案数量排名前三，具体数量分别为2329部、1214部和938部。而排名倒数前三的地区分别是青海、西藏和甘肃，具体数量分别为4部、13部和16部。

2.3.3 统计题材分布

如图7所示，从题材上来看，备案数量排名前三的题材分别是当代都市、近代革命和近代传奇，这三个题材的备案数量分别为3396部、1130部和709部。

3 结语

运用Python爬虫技术能够顺利的抓取所需数据。通过对数据的整理和分析，可以认为：从总量上来看，我国电视部备案数量整体呈稳步上升趋势；从地区分布上看，备案数量与地区经济的发达程度正相关，由于电视剧的拍摄和制作需要资本投入，发达地区拥有资本和人才优势，能够大批量的拍摄和制作电视剧；最后，从备案题材来看，当代都市题材的数量处于绝对领先地位，说明反映时代特征的当代题材剧最受资本和制作方的亲睐。

参考文献

[1]Yves Hilpisch.Python金融大数据分析[M].北京:人民邮电出版社,2015.

[2]吴剑兰.基于Python的新浪微博爬虫研究[J].无线互联科技,2015,(6):93-94.

Crawling and Analysis of Web Data Based on Python Crawler Technology

Xiong Chang

(Three Gorges Finance Limited Liability Company, Beijing 100038)

Abstract: Python crawler technology is simple and easy to use. Using Python language to write program to crawling the drama data on the SARFT's website. And we made a statistical analysis of the recorded data and draw the relevant conclusions.

Key Words: Python; Crawler; Data Analysis