基于网络爬虫的军事舆情态势挖掘研究

崔致远
(郑州外国语学校，河南郑州，450000)

摘要：随着网络技术发展，可利用的信息不断增多，而网络舆情就是可利用的信息资源之一。且我国军事问题一直是民众的焦点话题，如何较之以往更好的提取信息成为关键。在此背景下，本文基于网络爬虫和主题模型技术对相关网站的舆情信息进行挖掘研究。使用火车头软件对军事网站爬取，经过预处理后，运用Python语言实现对挖掘到的军事舆情数据的分析，从而实现对军事舆情提取精确、关键的信息，也为提取军事舆情信息提供了另一种较好的思路。

关键词：网络爬虫；军事舆情；数据挖掘

DOI:10.16589/j.cnki.cn11-3571/tn.2018.z2.068

0 引言

经过近几十年的发展，网络中数据量不断增大，其中可以利用的信息资源十分丰富。社交软件的不断涌现，使人们在社交平台上发表对事实的评论成为日常生活的一部分。在众多的舆情中，军事舆情是较为突出的一方面。网络中的军事舆情反映着对这类事件的民意导向，对我国军事决策十分重要。它对监控军事舆情，预测事件发展动向有不小的作用。因此，对网络军事舆情分析是不可或缺的。如何快捷、有效地从中获取信息，成为需要解决的一大问题。而网络爬虫是解决这类问题基础技术之一。

目前，国内基于 web 的舆情分析较少本，因此本文对军事舆情的研究选择以 web 舆情为基础展开。基于火车头采集器（网络爬取工具）对国内的某个军事网站页面进行爬取，得到近期国民较为关注的关键词。对其进行预处理后，基于 python 语言实现对数据的主题分析，整理为图样，数据表等形式。通过以上几步，实现对舆情的提取和分析。

1 研究现状

本文使用的关键技术之一是网络爬虫。网络爬虫，顾名思义，如同趴在巨大万维网里面的一只蜘蛛，它在盘枝错节的网络中自动漫游，抓取有用的资源，并加载到设立的数据库中。网络爬虫又称网络机器人。英文名有 Robot、Crawler、Spider 等。

在爬虫技术发展方面。国外的网络爬虫技术相对来说起步较早，已经有大量优秀作品。如斯坦福大学设计的 Google 的爬虫，康柏系统研究中心的 Allan Heydon 和 Marc Najork 设计的名叫 Mercator 的高性能爬行器，以及 Tnternet Archive、 UbiCrawler 等。当然国内也有如北大天网高性能网络爬虫。目前，网络爬虫的研究主要是围绕如何提高预测的准确性，降低计算复杂性，以及增加网络爬虫自适应性这几个方面展开。

在分析军事舆情的有关研究中，国内已有一些分析系统，如中科点击军犬网络舆情监控系统、北京拓尔思 TRS 络舆情监控系统等。但是，传统的分析方法中没有采用主题模型这种分析方法，使得在舆情分析中，难以十分精确的对数据进行提取分析，只是简单给出话题趋势变化。因此，使用传统的分析方法很难提取关键、精准的预警信息，而这将会影响到以网络舆情分析为基础的有关军事决策的准确性、适用性等。本文将实现以主题模型为方法的军事舆情分析。

2 Web 数据爬取

网络爬虫的工作机制如下：首先对待爬取页面进行检索，然后在爬取时有方向性的进行挖掘。数据爬取方向包含两种：第一种爬取方向是以横向宽度为爬取基准；第二种是以纵向的深度为基准。爬取完成后数据自动存储到新建立的存档文件中。

本文的 Web 数据爬取流程如下：

（1）找到可以爬取的相关军事网站，复制网站的地址，然后运用火车头软件，输入网站的链接即确认采集的网址规则。

（2）火车头软件根据提供的 URL（即网页网址）访问页面，查看网页的源代码，搜寻相关内容格式，以避免爬取到大量无用信息，使用软件中的标题，通过对网页中所需要的信息提取标识符来提取标识符中的文本。

（3）尝试对某个网址进行分析，检验准确性与可行性。当测试网页提取到的内容符合预期，运行程序。火车头软件将一个页面分析完成后，开始爬取队列中下一个网页，直到爬取结束。

（4）保存数据到数据库中并准备预处理，保存到。由于提取到的内容中会存在不必要的无用信息（如网页链接、字符、标点等等），需要进行预处理，运用查找、替换等方式，使最后的结果只保留需要的舆情关键词。

3 基于主题模型的军事舆情挖掘

■ 3.1 主题提取

基于提取到的数据，运用词频分析和主题模型来实现对资料的主题分析。主题模型技术是一种非监督的计算机的学习技术，用来分析文本中的隐含信息，其处理的文本通常信息量较大，较为复杂。它使用三层贝叶斯模型结构，即文档，

主题和词。在这三者之中，文档以及词是直接可以看到的，而主题是隐含在文档和词里，是需要进行分析的。同时，在文档和主题；主题和词之间是有关系的，是概率分布，而且概率的大小可以是有差别的。主题模型的思路是使每一个词以一定的概率选取到某些主题，并从中以一定概率选出某些词语。文档是由多个主题一起构成，而主题是词在一定概率上的分布，通过这种方法分析文档和词的潜在意义，即主题。本文将基于Python语言实现对挖掘到的军事舆情数据的分析。

本文的军事舆情挖掘结果如下，基于 Python 语言最终得到五个主题。如图 1 所示。分析可知，军事舆情数据的五个主题可概括为：中国军队新装备亮相以及世界维和；中国导弹部队；中国国际军事赛事承办和军事活动；中日和中美南海问题；国产新型军事装备及技术（下文中分别使用数字 0-4 表示）。

```
Topic #0:
中国 解放军 海军 战机 新型 部队 导弹 装备 国防部 印度 维和 专家 亮相 基地 台军 日本 世界 技术 直升机 军队
Topic #1:
官兵 陆军 导弹 集团军 台军 部队 军报 司令员 火箭 停止 训练 准备 台独 建设 军队 没有 安全 东风 重大 番号
Topic #2:
国际 军事 中国 比赛 大陆 中国空军 坦克 潜艇 进行 建军 周年 装备 我军 战略 国家 海军 赛事 承办 东海舰队 活动
Topic #3:
中国 编队 空军 钓鱼岛 海警 巡航 参加 辽宁 训练 台湾 飞机 海军 领海 举行 战机 专家 美国 军人 南海 我军
Topic #4:
回应 国防部 中国 航母 中方 军舰 辽宁 国产 发射 外交部 军售 无人机 海域 进入 台湾 军机 印军 飞行 北斗 武警
()
```

图 1　军事舆情主题挖掘结果

■ 3.2 主题相关性分析

进一步分析各主题间的关系可知，各主题在数据中出现的频率略有差别，如图 2 所示，主题 0 占的比重最大，1、2、4 所占比重相似，且主题 3 的频率最低。

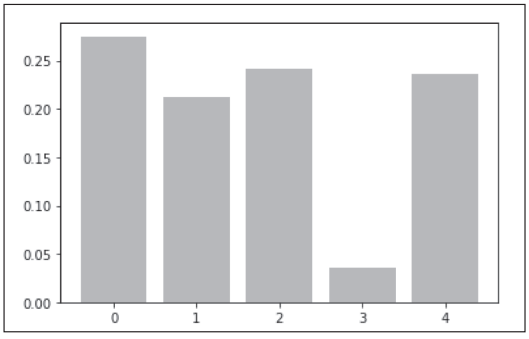

图 2　主题频率分布直方图

此外，基于不同主题之间的相关关系分析可知，不同主题间的相关关系有所差异，如图 3-4，散点对应主题间两两同时出现的概率。分析可知，由 0、1、2、4topic 得到的散点图是相似的，如图四显示。同时 0、1、2、4topic 与 3topic 的关系相似。

如图 3，横轴表示 0、1、2、4topic 中一个 topic 出现的概率，以及纵轴为此概率下的 3 topic 的概率。点的分布主要在横轴上，纵坐标的值多为零。由此可以分析出 3 topic 与其他 topic 的联系相对不那么紧密。如图 4，点的分布集中在等腰直角三角形的区域内，且各个位置都有分布。较之图 3，可以说是相对联系紧密的。

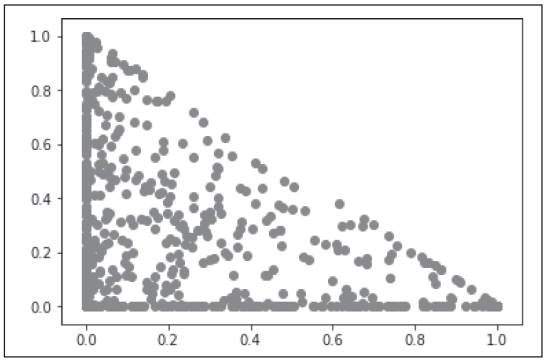

图 3　topic4 与 topic3 概率关联

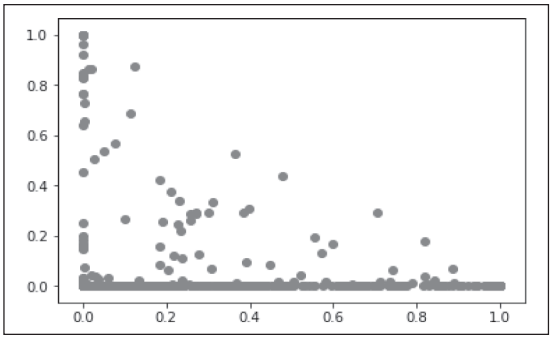

图 4　topic4 与 topic2 概率关联

■ 3.3 词云分析

基于主题提取结果绘制词云图，图中文字尺寸与词频正相关，可知 词语“中国”频率最高；“解放军”、“导弹”、“航母”、“海军”等词语次之。由此可推测 Web 军事舆情更倾向于基于这些关键词展开讨论，这也为舆情把控提供了基本方向。

4 结论

本文通过网络爬虫技术和 Python 语言实现对军事舆情数据的挖掘和分析。本文采用某个军事网站的国内军事舆情为挖掘资源，是用火车头软件提取，用 Python 语言进行主题模型分析，得出研究时段的网络军事舆情关注的焦点。同时，较之以往分析舆情趋势的方法，本文使用 Web 舆情数据基础作为数据源，基于主题模型军事焦点话题及其相关关系，可以更好的把握国民较为关注的军事事件，舆论导向，从中提取关键、精准的预警信息，从而为军事决策提供较为精确的舆情资料。同时，也为提取舆情信息提供了另一种较好的思路。未来将改进提取和分析方法，以进一步提高得到的舆情关键词的准确性和本文所用技术的实用性。

参考文献

＊ [1] 曾伟辉，李淼，曾伟辉 . 深层网络爬虫研究综述 [J]. 计算机系统应用，2008, 17(5):122-126.

（下转第 102 页）

不但涵盖了广泛的知识内容，同时也给操作人员的专业水平提出了严格的要求，设计人员必须具备丰厚的理论基础以及专业水平操作能力才能胜任。在这种背景下，带给传统电子产品设计过程的挑战更大了。智能化技术在电子产品设计中的应用可以有效地实现电子产品的设计，不仅减少了相应的人力物力资源，并同时显著提升了产品的生产效率，目前社会各界对于电子产品设计中的智能技术都非常关注。应用计算机网络技术不仅能够实现电子产品性能与功能的测试，并且用于研发的时间也大大降低，实现了设计环节的优化。整体来讲，电子工程行业未来的发展，已经无法离开现代智能技术。例如，在办公室中设计测试机器，一方面是带有一个传动带的 1×2 轴三角式机器人，另一方面是带有两个传动带和视觉系统的 2×3 轴三角式机器人。这两种技术的应用，有助于提升办公室检测工作的效率，优化工作质量。

■ 3.2 有效排除故障的产生

如今很多非人为的因素仍然会影响电子工程自动化控制系统的顺利运行，产生系统故障，所以系带加强处理自动化控制系统故障。然而控制系统故障使用传统的技术效果不佳，所以就需要找到更为科学有效的技术来应用，智能技术应运而生，大大帮助到排除电子工程自动化的系统故障。以过去在自动化控制系统中经常出现的故障情况来分析，通常出现的电子故障之间是有着共同点的，所以，智能技术中包括的神经网络、模糊逻辑、专家系统这些专业技术，都可以用来有效诊断系统故障，使我们在系统故障发生时更加及时和准备的发现问题所在，并第一时间找出对应的办法，从而实现帮助系统故障的排除，使系统可以正常运行。

■ 3.3 同时完成更多的操作任务

近年来由于智能技术的不断普及，对于电子工程领域带来了积极影响，其中应用智能技术有效实现了多元化的运行。智能技术具有可同时操作各类复杂任务，为系统正常运行提供保障，同时还不要求操作人员具有较高能力，能够允许操作人员同时进行多项系统的操作，另外因为应用了智能技术，实现了智能化控制，还降低操作任务的失误率，保证整个系统的正常运行的多重优点。智能化技术不但比传统操作系统可控制多个对象，并显著提升产品质量，给控制系统高速运行提供了可靠的基础。因此，智能技术的应用在电子工程领域的发展前景良好。例如，在制药业中应用注射器装载机，这一系统的机器人型号是 DSS—0800，最大轴数量是（5）XYZUA 三角式机器人，速度可以达到 400ppm，主要应用在制药行业中。这一设施可以在生产线中应用，用于放下产品的第五个轴，所有的伺服系统具备柔软性。设施在应用后非常容易实施清洁工作。

4 结论

不能否认，我国的经济技术正在逐渐快速发展，但是电子工程自动化智能技术领域仍有很多问题存在。智能技术近年来发展迅猛，智能技术作为一种现代科学技术已被广泛应用于许多领域。在电子工程自动化控制中，智能技术的应用在提高工作效率和确保信息的自动分类和收集方面非常有效，系统的稳定性在促进运行中起着非常重要的作用。通过分析智能技术在电子工程自动化控制中的技术特点，阐述了智能技术在电子工程自动化控制中的应用，提出了电子工程自动化控制中电子工程自动化控制的设计。在今后的发展中，本文为了解电子工程自动化技术系统提供了一定的理论借鉴。智能技术的应用正是体现出了与时俱进的精神，是社会进代、时代发展的需要。所以，进一步研究电子工程智能化智能技术非常必要。

参考文献

＊ [1] 刘继雷 . 智能技术在电子工程自动化控制中的应用 [J]. 电子技术与软件工程 ,2016(18):167-167.

＊ [2] 马永晟 , 于凤娇 , 何汉文 . 智能技术在电子工程自动化控制中的应用 [J]. 南方农机 ,2016(2):86-86.

＊ [3] 秦晓磊 . 智能技术在电子工程自动化控制中的应用 [J]. 数字技术与应用 ,2017(2):22-22.

＊ [4] 路海英 . 浅析智能技术在电子工程自动化控制中的应用 [J]. 电子制作 ,2016(18):54-54.

＊ [5] 崔凯 . 浅析智能技术在电子工程自动化控制中的应用 [J]. 装饰装修天地 ,2017(12).

＊ [6] 谢欢 , 宋培 . 智能技术在电子工程自动化控制中的应用探讨 [J]. 工程技术 : 引文版 ,2016(12):00270-00270.

（上接第 143 页）

＊ [2] 胡晟 . 基于网络爬虫的 Web 挖掘应用 [J]. 软件 , 2012(7):145-147.

＊ [3] 王丹 . 基于主题模型的用户画像提取算法研究 [D]. 北京工业大学 , 2016.

＊ [4] 刘晓亮 . 基于维基百科的军事舆情论坛话题追踪方法 [J]. 计算机应用 , 2012, 32(11):3026-3029.

＊ [5] 杨旭东 . 网络舆情监控系统关键技术研究 [J]. 信息网络安全 , 2016(9):251-256.

＊ [6] 李晨 , 杨子江 , 朱世伟 , 等 . 基于 Hadoop 的网络舆情监控平台设计与实现 [J]. 计算机技术与发展 , 2016(2):144-149.

＊ [7] 周红福 , 贾璐 , 张婷婷 , 等 . 微博舆情分析中信息转发路径提取方法研究 [J]. 信息网络安全 , 2016(4):61-68.

＊ [8] 王曰芬 , 杭伟梁 , 丁洁 . 微博舆情社会网络关键节点识别与应用研究 [J]. 情报资料工作 , 2016(3).