Advances in Social Science, Education and Humanities Research, volume 582
Proceedings of the 2021 6th International Conference on Modern Management and Education Technology (MMET 2021)

Big Data Web Crawler Analysis of Online Professional Course Requirements

Song Shuo[1,a], Wu Guangcuan[2], Peng Yiping[1]

[1]*College of Science, Wuhan University of Science and Technology, Wuhan 430065, China*
[2]*Guangdong Hongtu Wuhan Die Casting Co., Wuhan 430200, China*
[a] *songshuo@wust.edu.cn*

ABSTRACT
In the context of the era of big data, the development and progress of network technology has brought learners multiple learning methods and strategies of personalized learning choices.Although the development of online courses in our country has begun to take shape, many problems have emerged in the implementation process. There are still obvious obstacles in the design of courses, the use of teaching tools, and the transformation of educational concepts. As for professional courses, due to their own characteristics, the online teaching of professional courses is particularly hard. In view of these problems, this article has done the following work: Use Python to crawl online comments on professional courses, generate text documents and then perform text analysis. In the text analysis, Python jieba library is first used for word segmentation and word frequency statistics, and then Python wordcloud library is used to generate comment wordcloud to study learners' demands for online professional courses. On the basis of the above analysis, summary and suggestions are given to provide reference for the development of online professional courses in our country, to promote the good development of online professional courses, and to provide convenience for relevant professional learners.

Keywords: *Online professional courses, Web Crawler, Text Analysis.*

在线专业课程需求的大数据网络爬虫分析

宋硕[1, a] 吴广川[2] 彭奕萍[1]

[1] *武汉科技大学理学院，湖北，武汉，430065*
[2] *广东鸿图武汉压铸有限公司，湖北，武汉，430200*
[a] *songshuo@wust.edu.cn*

摘要
目前正处在大数据时代背景之下，互联网及相关产品的发展给学习者提供了多重学习方式和个性化学习的选择。我国的在线课程发展虽已初见规模，但在实施过程中涌现出许多问题：在课程的设计上、教学工具的使用上、教育观念的转变上还存在着明显的障碍。专业课程又有其自身的特点，所以专业课的在线教学问题尤为突出。鉴于这些问题，本文做了以下工作：用 Python 爬取网上对于专业课程的评论，生成文本文档再进行文本分析。进行文本分析时先用 Python jieba 库对文本做分词处理并统计词频，再利用 Python wordcloud 库生成评论词云，研究学习者对在线专业课程的需求。并在以上分析的基础上给出总结和建议，为我国在线专业课程的发展提供参考，以期促进在线专业课程良好发展，为相关专业学习者提供便利。

关键词： *在线专业课程； 网络爬虫； 文本分析*

1. 前言

随着大数据时代[1]的来临，大数据时代背景使得在线学习专业课程成为可能，网络技术不断发展，为在线专业课程提供了有力的技术支持。同时，在线课程能够改变课程的开发主体、丰富高校课程内容、改变高校课程载体和课程实施手段，有助于提升教师的专业知识和技能，提高课堂教学效果，有助于培养学的主动学习能力，丰富学生的学习生活。这些都是在线专业课程发展的机遇所在。

国内外学者对在线教育进行了不同角度和层次的研究[2]。目前，我国在线教育还处在探索阶段，学习者视角下在线课程开发与设计方面的相关研究还不成熟，研究成果较少。本文将从“用户对在线专业课程的需求出发”，坚持学习者为本，从学习者的角度出发研究需求，为我国在线教育发展提供建议。本文希望通过大数据爬虫技术，调查大众对在线专业教育课程的需求和评价，了解国内在线专业课程现状，分析我国在线专业课程未来发展的机遇与挑战，借鉴国外相关经验，为我国在线专业课程的发展提供参考，以促进在线专业课程良好发展，为相关专业学习者提供便利。

2. 基于文本挖掘的在线专业课程需求调查

由于传统问卷调查在受调查人群覆盖上的局限性，很难覆盖到各个不同年龄、专业的使用者。所以仅凭调查问卷的结果来推断总体可能有的特征是不够严谨的，缺乏一定的合理性。因此，在这里，我们运用网络爬虫的方法，从网上获取相关文本信息，并利用 Python 做简单的文本分析，进而从中提取出用户对在线专业课程的关注点来找出用户的需求。

由于很多在线专业课程的投放大多是在一些较为著名的平台。考虑到这点，本文主要运用 Python 语言爬取投放在 MOOC 平台上的一些精品课程的评论[3]。由于 MOOC 平台上并非所有的课程都是专业课程，再加上平台上的课程投放量之多，难以爬取到所有课程的评论。所以本文从 MOOC 平台上的课程分类出发，爬选取具有代表性的一些专业课程的评价数据，我们选取了计算机类、经济学类、理学类、文史类和工科类共五个分类下的最热门的三门专业课程，共十五们课程的评价。之后再利用 Python 进行中文分词、生成词云，根据结果从中分析提取课程学习者的关注点。

2.1. 数据的爬取

进入中国大学 MOOC 的网站，从首页的导航栏看到课程的分类。以计算机课程类别下的热门课程《Python 语言程序设计》为例，课程界面如图 1 所示。通过对课程页面的简单操作后发现，该网页是动态网页。不同于静态网页，动态网页可以动态解析 URL 中参数的变化，关联数据库并动态呈现不同的页面内容。由于需要爬取的网页不再是一个简单的 HTML，所以选择使用模拟浏览器运行的方式来实现在浏览器中看到是什么样的抓取的源码就是什么样[4]。

先进行爬取课程评论的准备工作，选择使用 Selenium 库来模拟浏览器运行。Selenium 是一个自动化测试工具，可以驱动浏览器执行特定的动作，如点击、下拉等操作，同时还能获取浏览器当前呈现页面的源码。使用 Selenium 库时要先声明浏览器对象，用代码 driver = webdriver.Chrome( )来完成浏览器对象的初始化并将其复制为 driver 对象，然后再使用 get( ) 方法访问页面。对于课程评论的爬取，从课程主页看，需要使用 Selenium 驱动浏览器完成模拟点击评论，并在一页二十条评论爬取完成后继续模拟点击下一页操作，待页面跳转后继续爬取当前页的评论，这样循环下去直到将所有的评论都爬取下来[5]。将课程评论都爬取下来后，为了方便后面的文本分析，将爬取下来的评论写入 txt 文档中。本次爬虫爬取了计算机、经济学、理学、文史类和工科类这五个不同类别下的 15 门热门课程的评论，共计有 16840 条评论。将这些评论按照课程类别分别写入五个不同的 txt 文件内。

图 1 课程界面图

2.2. 爬取的评论绘制成词云

想要从爬取到的评论中获取信息，则需要进行文本分析，在这里选择将大段的文本进行中文分词处理，再绘制词云将分词结果直观的呈现出来，以方便进行下一步的用户需求分析。在分词时选择的是 Python 里的 jieba 库来对文本进行分词，分词后发现有结果中含有大量类似于“的”、“我”、“并且”等对分析无意义的词，还会影响后面词云生成的效果。所以选择导入停用词表，本文选择的是从网上下载到的哈工大的停用词表扩展版。导入停用词表后，去除了大量的无意义词和字符，分词结果变得更加清晰也更加有使用价值，也方便了后面的词云生成过程。

使用 Python 中的 wordcloud 库来绘制词云让词语的频率变得可视化，更加清晰直观，一般来说出现次数越多的词语在词云中占的面积就越大。绘制词云时先选取好背景图将背景图路径导入，选取好中文字体，再根据每个文档分词的结果设置最大显示的词量、最大字号，最后将生成的图片保存起来。生成的词云如下图 2 到图 6 所示。

2.3. 结果分析

在词云图中，所占面积越大说明该词在课程的评论中出现的频率越大。而在评论中出现的频率可以在一定程度上反映学习者在这个点上的关注度，所以可以通过对词频的统计，来总结学习者对在线专业课程的需求。

从计算机类课程的词云图 2 中可以看到出现频率较高的词有“老师”、“讲课”、“清晰”、“易懂”、“实习课”等。可以感觉到，学习者比较关注老师的讲解是否清晰易懂，课程内容安排是否合理，内容是否循序渐进、易于学习和理解。除此之还，还可以看到类如“实习课”、“习题课”等内容。由于计算机课程有很多都是与编程和实践有关的，这要求学生有一定的编程练习，要有足够的实操训练和代码积累量才能学好一门编程语言，所以计算机类的课程评论中出现了不少关于“实习课”和“习题课”这些词语。这也告诉我们，由于计算机类课程的特殊性，计算机类课程可以考虑改善学习者关于实习课和习题课的体验。比如对于编程语言课程来说，可以考虑搭建云端 IDE，学生使用账号登入后就可以在上面写代码编译运行，并将代码作为作业提交，极大地方便学生。还可以选择针对较难的习题，录制点拨视频发布在课程平台，帮助学生寻找解题思路，以提升学生对于习题课的体验。

图 2 计算机类课程评论词云

图 3 经济学类课程评论词云

从经济学类课程的词云图 3 中可看到，学习者关注最多的还是老师的讲解是不是够清晰，是否能够深入浅出，是不是足够通俗易懂。同样学习者对于课程内容也是很关注的，课程内容在评论中出现的频率也是很高的。同时在经济学类课程评论的词云中发现，“生活”、“例子”这两个词出现的频率也挺高的。由于经济学课程大多与现实生活联系的较为紧密，有很多经济学的概念也只有跟实际生活联系起来才能够更好的理解。所以经济学类的课程，无论是在课程内容设置上，还是在老师讲授的过程中，最好都能够做到更多地更好地引进实际生活中的例子，来帮助学生学习经济学相关内容。

从理学类课程的词云图 4 中可知道，老师是在评论中出现最多的词语了，学习者比较在乎老师的讲解是否清晰。相比于其他类别的课程，学习者好像并不是很在意课程的内容，这可能是由于理学类的课程概念相较于其他课程更加的抽象，所以会更加的依赖于老师的讲解，导致学习者对老师的需求会更多相对弱化了对课程内容的需求。

图 4 理学类课程评论词云

图 5 文史类课程评论词云

从文史类课程评论的词云图 5 里可以看出，学习者对于在线专业课程的需求主要还是集中在对于课程和老师的需求，学习者希望老师的讲解够浅显易懂，课程内容设置合理能够让人受益匪浅。

图 6 工科类课程评论词云

在工科类课程的评论词云图 6 中可以看到，首先在学习者的评论中出现最多的还是老师这个词，说明学习者非常看重老师在一门课程学习中发挥的作用。他们希望老师有很好的教学能力，能够深入浅出地讲解知识点。还有就是课程的内容也很被看重，所以在课程内容的设置上，要尽量做到合理。课程内容设置有能够激发学习兴趣，循序渐进难度设置要合理。在工科类课程的评论词云中还出现了教材这个词，说明工科类课程的学习者对于教材是有更多的需求的，所以在工科类课程的设置中可以考虑在课件上下更多的功夫，提供课件和教材的下载渠道，供学生下载。

根据生成的词云可以看出，不论是哪个类别课程下的评论，“老师”和“课程”在评论中出现的次数都很多。这说明学习者对于在线专业课程的授课老师和课程内容比较在意。从生成的词云图上还可以看到，类似于“浅显易懂”和“清晰”等对课程内容和老师授课的描述，也进一步说明学习者对于一门课程的评价多是从课程内容和授课老师出发考虑的，也就是说课程内容和授课老师的质量在一定程度上会影响到学习者对于课程的评价。

3. 结论与建议

基于对爬虫所获取到的数据的分析，得到如下结论与建议：

（1）课程内容精品化是亮点

从对爬虫得到的数据分析后可发现学习者对于在线专业课程内容的重视程度很高。所以对于专业课程内容安排上要精品化，提高在线专业课程的制作门槛，提升精品课程在在线专业课程的占比。并且对于不同类别的专业课，应根据专业课本身的特点有所侧重地去开发重难点的讲解视频。

（2）授课老师的专业化是关键

从爬虫获取到的数据的分析，可以看出老师对学习者在线学习的体验有很大影响。课程的授课老师要足够专业，要有足够的专业知识，这样才能使得学生信服老师而产生信任感从而愿意跟着老师进行学习。这一点其实与线下课程也是一样的。

（3）与学生有良好的线上互动非常重要

由于不同于传统课堂学习，在线专业课程的学习实在线上进行的，学生在无监督状态下容易走神，所以老师应该学习一些线上互动方法，开通和学生的互动渠道，使得学生能够集中注意力完整地上完每堂课。

项目基金

本论文由湖北省教育厅人文社会科学青年研究项目《大数据时代背景下在线专业课程需求的调查分析研究》（编号 18Q030）和武汉科技大学校教学研究项目《基于大数据分析的统计专业在线课程建设问题及改进策略研究》（编号 2019X054）资助。

REFERENCES

[1] Viktor, M.-S., Kenneth, C., Sheng, Y.y., Zhou, T., Trans. (2013) Big data: A revolution that will transform how we live, work, and think. Zhejiang people’s Publishing House, Hangzhou.

[2] Guan J., (2014) China's online education development status, trends and experience reference. China Audio-visual Education, 331：64~66.

[3] Shi, X.C.,Li M.L., (2016) International MOOC research hotspots and trends. Open Education Research, 22(1)：93~97.

[4] Cui, Q.C. (2018) Python3 Web crawler development combat. People's Posts and Telecommunications Press, Beijing.

[5] Luo, G. (2017) Full analysis of Web crawlers. Electronic Industry Press, Beijing.