DOI:10.16298/j.cnki.1004-3667.2020.04.05

大学生在线学习体验的聚类分析研究*

贾文军 郭玉婷 赵泽宁

摘 要: 为深入了解疫情背景下大学生的线上学习体验,给未来线上教学的开展提供相关参考,利用爬虫软件,搜集微博平台上关于大学生网课的评论,使用SPSS软件和Python编程对收集到的文本数据进行分词,词频统计和聚类分析。根据评论的类型,分析在线教学中学生体验的现状:课前存在硬件设施、网络环境和学校支持不到位的情况;课中的问题主要集中于教学平台不统一,授课现场组织不到位,学生学习状态不在线;课后的作业任务重,学生虽然能够接受在线教学这种模式,但更期待能返校上课。

关键词: 线上教学;学生体验;网课评论;聚类分析

一、引言

国内外学者关于大学生的学习体验研究,大都采用问卷调查或者以质性访谈和理论分析的形式进行。考虑到我国此次线上教学的大规模特征,若通过人工设计问卷进行调查,难以真正发现学生的内心感受,且在操作过程中无法保证其真实性,本研究通过"爬虫",利用社交平台上的文本数据,对疫情背景下大学生的学习体验进行了聚类分析。该方法不同于传统的问卷调查数据和访谈资料,可以更真实地表达学生感受[1]。正是基于"真实性"的思考,本研究利用国内微博这个社交平台,采集了学生关于在线教学评论的文本数据,并运用聚类分析(clustering analysis)作为数据挖掘方法,以此来识别和刻画在线学习者的学习体验[2]。在数据收集的基础上,通过数据去重、特征提取、聚类过程、聚类结果评估五步骤,对当下大学生的在线学习体验进行了基本分析。

本研究采用质性研究方法,以微博文本数据作为数据来源,以学生为中心,运用网络爬虫搜集微博中高校大学生关于网课体验的评论与话题,通过文本数据聚类分析深入挖掘文本数据,呈现出特殊时期学生在课前、课中、课后真实的学习体验状况,分析影响学生体验的关键因素。

二、研究方法与工具

本研究采集微博数据的工具是网络爬虫,能够根据目标需求爬取相关网页上的信息。网络爬虫抓取的微博数据来源分为三类:第一类是全国2 688所高校的微博主页搜索与网课相关的博文,搜集每条博文下的相关评论;第二类是热点话题如"大学生网课""大学生网课日常""当代大学生网课现状"等的相关评论;第三类是来自于各教学平台使用评论,使用的平台有钉钉、中国大学MOOC平台、QQ直播、腾讯会议、腾讯课堂、ZOOM、雨课堂、学习通、智慧树等。爬取出来的文本数据以Excel文件的形式呈现。数据采集示例如表1和表2所示,采集的时段为2020

表1 博文评论数据搜集示例

博主头像	博主ID	博主昵称	博主主页	评论内容	发布时间	点赞数	回复数	抓取时间
https://tvax2.sinaimg.cn…	5145789931	lio_one	https://weibo.com/5145789931?……	云课堂又叫职教云。垃圾中的战斗机看完的课件直接给我清零	2月28日12:13	0	0	2020-03-12 16:58:49.0

表2 话题讨论数据搜集示例

话题	博主昵称	博主ID	博主主页	博文	博文网址	发布时间	发布终端	转发数	评论数	点赞数
#大学生网课日常#	搞笑南叔	5516030215	https://weibo.com/5516……	#大学生网课日常#上网课的女大学生和电脑精来了!!着实是被网课逼疯了	https://weibo.com/551……	3月09日12:52	电脑	225	490	8 327

* 本文系2019年度国家社会科学基金教育学重点项目"中国特色、世界水平的一流本科教育建设标准与建设机制研究"(AIA190014)的研究成果

年1月29日至3月17日，数据总量为12 058条评论。

整理爬虫搜集的文本数据作为基础数据，利用Python编程中Jieba分词工具包将文本进行分词处理。Jieba中文分词可以将句子最精确地切开，适合文本分析，把句子中所有可以成词的词语都扫描出来，通过分词后得到词频统计数据。根据词频数据生成词频矩阵，利用SPSS软件中的系统聚类分析工具生成系统谱系图，以便直观地分析评论的类型。在聚类分析基础上，再运用文本分类中用于特征选择的卡方统计来计算关键词与类别之间的关联度，即CHI值。CHI值越大，对应特征项所包含的与类别相关的鉴别信息就越多。将词频数排名前100名的词语进一步聚类，得到词群后对其类型定义，最后得到了评论的具体类型，再依据课前、课中、课后分析疫情期间在线教学中学生学习体验的现状。

三、数据分析

（一）数据去重

将12 058条博文评论内容的文本进行数据“清洗”。去除话题评论中所含有的共同话题词，如＃大学生网课＃、＃大学生网课日常＃等，避免这些主题词的词语重复计数到关键词的词频中而影响后续的分词和词频的统计工作；再剔除评论中出现的英文字母或异常词及无效文本评论。整理完毕后将Excel文件转换成Txt文件使用。

（二）数据处理

将清洗后的数据放入Python程序的词云生成器中，先对文本数据中的所有词频进行词频统计，之后提取词频数排行前100的词语列表，示例显示排行前20的词语见表3，并生成词云如图1所示，词语的字体越大说明它的词频数越高。

然后将文本数据导入Python软件生成词频矩阵，部分词频矩阵如表4所示。词频矩阵中每行代表一条评论的内容，每列代表一个关键词，矩阵中的1表示此关键词是相应评论中的关键词，而0代表的是这条评论中不含对应的关键词。

表3 词频数排行前20词频表

排序	词语	词频数	排序	词语	词频数
1	课堂	2141	11	姿势	388
2	视频	1169	12	手机	373
3	上课	1067	13	课程	365
4	上网	1014	14	在线	358
5	作业	929	15	时间	344
6	开学	756	16	在家	344
7	会议	435	17	网页	303
8	直播	430	18	超星	265
9	软件	409	19	电脑	232
10	签到	394	20	回学校	197

图1 词频数TOP100词语生成的词云

表4 部分词频矩阵

	网课	学习	大学生	…
评论1	0	0	0	…
评论2	1	1	0	…
评论3	1	0	1	…
…	…	…	…	…

（三）聚类分析

文本数据通过上述步骤转化成了0-1值的数据矩阵，将上述词频矩阵导入SPSS进行系统聚类分析。系统聚类分析选择聚类效果较好的组内联结法，此法采用简单匹配系数度量评论之间的相似性。简单匹配系数是当两条评论在关键词上的数值相同时出现的频率，频率越高说明两条评论越相似。经过系统聚类后的谱系如图2所示，可以看出，根据目前采集到的文本数据，大学生网课评论的文本可分为7类，选择距离阈值为22.5，聚类统计结果如表5所示。

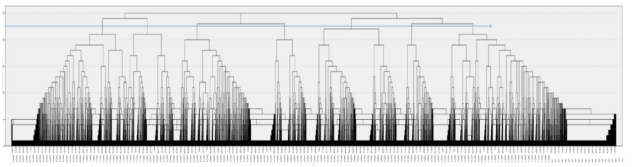

图2 系统聚类分析谱系图

表5 聚类结果

类别	第一类	第二类	第三类	第四类	第五类	第六类	第七类
评论数量/条	637	2 473	1 408	1 740	1 322	789	664

（四）基于卡方统计的聚类结果

利用Python编程来计算卡方统计的聚类结果如表6所示。每一类都是一个词群，我们针对每个词群进行类型定义。

第一类的高度关联词突出表现为回学校、打卡、不想、上学、学校，故定义为大学生的“网课意愿”；第二类关联词与教学平台相关定义为“平台体验”；第三类关联词主要围绕学生在线听课现场的状况，可定义为“授课现场”；第四类关联词则围绕学生的听课设备，将第四类关定义为“个人设备”；第五类关联词涉及学生的课堂视频学习以及课后作业布置，故

表6 高关联度关键词及类型定义

类别	高关联度关键词	类型定义
第一类	回学校,上课,快乐,为啥,打卡,语音,天天,家里,不想,学校,下课,不用,授课,上学,晚上	网课意愿
第二类	网页,学习,在线,课程,超星,签到,答案,进不去,告诉,垃圾,打开,体育课,教学,软件,不停	平台体验
第三类	关麦,可爱,听到,老师,同学,点名,回答,声音,提问,讲课,群里,直播,听课,屏幕,不好	授课现场
第四类	课堂,在家,慕课,校园,不用,学校,学生,摄像头,时间,系统,电脑,手机,疫情,开课,微信	个人设备
第五类	视频,眼睛,作业,布置,喜欢,提交,希望,课后,记录,看着,不到,体育课,家里,现场,结束	课业任务
第六类	姿势,上学,明白,上网,状态,不想,校园,通知,分享,好多,快乐,眼睛,欺负,笔记,翻车	学习状态
第七类	开学,啥时候,感觉,东西,笔记,线上,不想,希望,天天,在家,一点,好多,教学,结束,学院	期待开学

将第五类定义为“课业任务”;第六类关联词定义为“学习状态”;第七类关联词则定义为“期待开学”。

根据以上数据分析结果,通过搜集到的12 058条微博评论数据,将聚类分析出的七类结果做成雷达图(见图3)可以看出,微博评论关键词中有关于网课平台体验(2 473条)的描述最多,此类别高关联度关键词占所有评论1/4的比例;个人设备(1 740条)的评论次之,此类别的高关联度关键词约占所有评论的14.4%;授课现场(1 408条)和课业任务(1 322条)两类别的高关联度关键词所占比例相近,各占到评论数据的1/10左右;网课意愿(637条)、学习状态(789条)和期待开学(664条)三个类别的高关联度关键词所占比例最少,都在1/20左右。

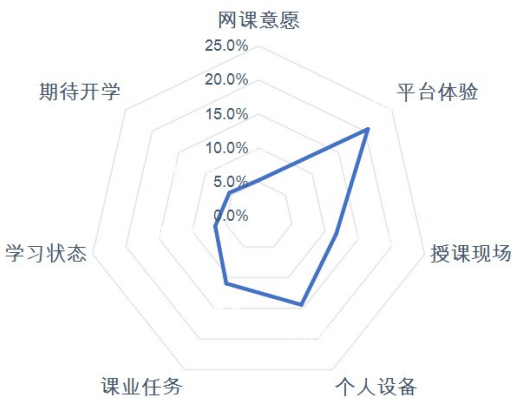

图3 评论类别百分比雷达图

四、研究发现与讨论

根据聚类分析的类别图,对于所搜集到的评论数据,本研究基于学生体验的立场从课前、课中和课后3个维度对线上教学进行讨论。

(一)课前

对于线上教学课前的分析主要考察学生在课前是否有相应的线上教学外部硬件设备,以及学校是否提供相关的支持服务。学生的线上教学外部硬件设备包括手机、电脑、平板等电子设备,同时也包括地区的网络基础设施,即学生所在归属地的网络建设状况和家庭的网络环境。学校提供的支持服务包括对于贫困学生的流量补助、课前教师的线上教学培训和教学秘书的配备,以及学校在教学安排前是否了解每一位学生所在地区和家庭的网络状况,对于线上教学平台的选择是否有统一的规划等。

1. 外部硬件设备。在第四类评论中,高关联度的关键词有电脑、手机、微信、摄像头、系统、时间、微信等,研究发现线上教学中外部硬件设备方面的问题在很大程度上影响到学生的学习体验。在评论中,有一部分网友因为处在信号不稳定的农村偏远地区或是家庭经济条件落后,存在着电子设备数量稀少、性能不足、家中没有无线网等现实问题,往往导致电子设备无法承载相关教学软件、手机流量不足和无法正常上课等问题。

2. 学校支持服务。在对于贫困学生的流量补助方面,有不少学校为确保学生的学习不被经济因素所影响,对家庭困难生发放了一定金额的临时困难补助和学习流量补助,如长沙理工大学、河北大学和西南民族大学等,因此也有不少学生在各大高校的微博平台下表示对学校的感谢。

为应对疫情,也有不少高校在线上教学正式开课前对授课教师展开线上教学的培训工作,但评论数据中发现仍有许多学生反映教师由于对平台功能了解不足,操作不熟练,进而导致上课期间,教师在平台操作上花费了大量的时间精力,导致课程时间延长、占据学生大量幕前时间和教学时间有效性低等问题。再加上学校未统一规定课程教师所使用的线上教学平台,造成学生所选的课程中需要使用到的线上平台软件繁多,学生需一一下载注册并加入不同的课程班级群。同时由于学校未制定出一份与教学平台相关的课表,许多课程也未配备教学秘书,导致学生时常弄混每门课程的上课平台、教学安排、作业布置、作业提交等问题。如网友提到的“老师讲了大半节课才发现没有开画面”“一个上午过去了,什么都没有学到,就看老师在那里调试软件了”。

另外研究发现在我国新疆、西藏或国外地区的学生与北京时间存在着一定的时差。学校需要提前了解这些地区学生的状况,并在进行课程时间安排时考虑这些地区的时差问题。

(二)课中

对于线上教学课中的分析主要考察上课过程中平台的运行状况、授课现场情况和学生学习状态。

平台的运行情况包括是否有卡顿、崩坏、无法打开等问题，以及平台页面的设计和平台功能便利性。授课现场情况包括课堂秩序的维持、课程流畅性和师生之间的互动。学生的学习状态包括学生在学习过程中的听课姿势、主动参与课堂情况和学习认真程度。

1. 平台运行状况。在线上教学开展以来，#学习通崩了#和#慕课崩了#等话题就频上热搜，在“平台体验”类别中，高关联度的关键词有“进不去”“垃圾”“打开”“签到”等，研究发现微博评论中学生对于平台体验发表了最多的评论，平台体验是影响学生体验的重要因素之一。疫情的突然爆发，导致线上教学平台承载客户量激增，在运行过程中产生了不同程度的卡顿、闪退，甚至是平台彻底崩溃。有网友表示“崩了！因技术问题导致逃课”“没啥体验，因为我根本进不去”。

在平台页面设计和功能便利上，评论数据发现课程授课方式常常是根据平台的功能进行设计。大部分软件都具有可以录制上课视频的功能，学生可以在课后进行回看，有网友表示“老师上课讲的没听到的可以回放再听，非常开心”。部分学校的老师使用的是提前录播好的课程，在关于录播课程的评论中学生有不同的感受，有网友认为“录播可以倍速，但是有时候做笔记要暂停”，同时还有网友指出个别教师课程不进行直播，录播课程还使用的是别校视频。线下的课程点名在线上教学中变成了课程签到打卡，一般在上课前教师会要求该门课程的学生在平台软件上进行签到打卡，并计入平时成绩，在前一百的词频“签到”位列第十名，也可看出学生们对于“签到”的评论量较多。有不少网友对平台的签到打卡功能使用的便利性表示怀疑，而打卡功能有时出现问题，导致打卡不成功，造成学生来上课了而“缺课”等尴尬局面。

2. 授课现场状况。结合第三类评论的高关联度关键字“关麦”“点名”“提问”“直播”等，在搜索评论数据时发现，授课时很多同学在上课的过程中会忘记关麦，进而在教师直播时会出现杂音，而一些老师在授课的过程中，老师所处的外部环境中的噪音也会影响到课堂秩序，学生往往会被额外出现的声音所吸引。由于师生互动也要通过“连麦”的方式，连麦的延迟性也使得教师在组织学生进行课堂讨论时出现杂乱无序的现象。还有不少网友对于教师的点名存在着惧怕心理，如“连麦好绝望”“点名有点怕”“抽人答题让我很不快乐”。但研究发现也有不少教师在进行线上教学时想出了别具一致的方式与学生进行互动，如课前课后播放铃声提醒，课上发放红包点名和与同学们说晚安等，获得了一致好评。

在课程流畅性方面，造成课程无法流畅进行的原因有网络、平台性能稳定问题，也有教师和学生的个人问题。

3. 学生的学习状态。学生的学习状态决定了学生的学习效率和学习质量。疫情期间，学生的学习环境仅局限于家中一隅天地，无法体验到与学校所营造出的相同的学习情境性。有不少网友在上课的过程中存在着“姿势”问题，而“姿势”在词频中排名11，可见，在线上教学中学生的听课姿势与线下教学存在着很大的差异。如“刚开始是端正地坐着，上着上着背就驼下去了，然后就开始站起来边甩腿边看了”。在家中进行学习不少学生感到了自律和态度的重要性。如“其实上网课学的多，看不完的还能课后看，大学生活好充实”“地点不重要，重要的是我们的态度”。但也有网友提到在家中学习只有一个人，有学习的孤独感，并且没老师监管很不自觉。

（三）课后

1. 课业任务。老师布置的课业任务是教学中必不可少的一个环节，学生不仅可以通过课后作业巩固知识，同时也是检验教师教学效果的一块试金石。由于教学形式的改变，课程效果的不确定性使得教师为了学生能够在新的教学形式下达到与线下教学相同的教学效果，增加了课后作业的任务量，但也无形之中给学生增加了学业压力，同时由于在家中学习没有相应的教材支持，很多学生对于知识点的理解存在困难。如网友对此评论“网课安排的那么密，还天天布置那么多作业，我怎么可能写完啊？”“没有课本，只是吸收知识，根本就不消化”。由此可见，视频只是辅助学习的一种手段，还需要学生通过书籍阅读进行思考才能产生更好的学习效果。

学生在完成作业后，需要到各个平台去提交自己的作业。由于网络平台的硬件存储能力，学生所处的网络环境以及自身的设备问题，提交作业过程往往繁琐且耗时过长。另外在上课和完成作业的过程中，长时间面对电子屏幕造成的视觉疲劳成了学生们评论的焦点。

令人欣喜的是大家对体育课的评价，称之为“最自由的课堂”。区别于传统的体育课，线上的体育课教学给学生提供了更多样化的教学方式。有的老师用Keep软件和学生一起健身，这种别开生面的形式给体育课的学习增加了不少乐趣。

2. 网课意愿。在聚类分析的“网课意愿”和“期待开学”类别关键词中，出现“回学校”“不想”“开学”“啥时候”“结束”等关键词，而“开学”和“回学校”在前20的词频分布中分别位于第6位和第20位。线上教学流程的繁复性、网络的不稳定性和一些外部因素使得学生在线上学习中所付出的时间和精力较线下教学呈翻倍的变化，同时对于学生的心理状态和自我管理也造成了不小的挑战。大家也纷纷表示出更怀念与老师面对面上课和赶紧回学校上学的殷切希望。虽然在网络评论中学生对于网课意愿不强，但同时学生也关注到教师为线上教学的付出，有网友评论说“我觉得老师们为了直播挺辛苦的，感谢他们”。

五、结语

本研究关注疫情期间大学生的在线学习体验，以网络爬虫所获得的关于学生网课体验的微博文本数据作为研究基础，通过文本聚类分析得出大学生学习体验主要可分为七大类别，依据这七大类别中的高度关联词及评论数据分析学生在课前、课中、课后的学习体验现状。其中课前这一层面包含个人设备问题；课中这一层面包含平台运行状况、授课现场情况和学生学习状态；课后这一层面包含课业任务、网课意愿和期待开学。根据学生体验现状，线上教学并不能完全取代线下教学。未来的线上线下教学之争，必然包括学生体验之争。本研究结果为未来线上教学的开展提供了有价值的参考，也给当下线上教学的相关工作者提供了真实的现状反馈。

（谢作栩教授为本文写作提供了精心指导，特此感谢。）

（贾文军，厦门大学教师发展中心博士研究生，福建厦门 361005；郭玉婷，厦门大学教师发展中心研究助理，福建厦门 361005；赵泽宁，厦门大学教师发展中心研究助理，福建厦门 361005）

参考文献

[1] ARORA S, GOEL M, SABITHA A S, et al. Learner groups in massive open online courses[J]. American Journal of Distance Education, 2017, 31(2), 80-97.

[2] CABEDO R, EDMUNDO T C, CASTRO M. (2016). A Benchmarking Study of Clustering Techniques Applied to a Set of Characteristics of MOOC Participants. 2016 ASEE Annual Conference & Exposition,New Orleans, Louisiana.

Clustering Analysis of College Students’ Online Learning Experience

JIA Wenjun GUO Yuting ZHAO Zening

(*Xiamen university , Xiamen 361005*)

Abstract: In order to deeply understand the online learning experience of college students and provide relevant references for the development of future online teaching, we used crawler software to collect comments on college students’ online courses on the Weibo platform, and used SPSS software and Python programming to perform word segmentation, word frequency statistics, and clustering analysis on the collected text data. According to the type of comments, we analyze the current status of student experience in online teaching: there are hardware facilities, network environments, and inadequate school support before the class; the problems in the class are mainly concentrated on the inconsistency of the teaching platform, the organization of the teaching site is not in place, and the students’ learning status is not online; homework assignments after class are heavy. Although college students can accept the form of online teaching, they are still looking forward to returning to school.

Key words: online teaching; student experience; online course review; clustering analysis