2018 3rd International Conference on Education, Management and Systems Engineering (EMSE 2018)
ISBN: 978-1-60595-591-9

Online Behavior Analyses of Undergraduates from University B: From the Perspective of Big Data

Ke-sheng LIU[1,2], Meng-xing WANG[2,3] and Yi-kun NI[1,2,*]

[1]School of Economics and Management, Beihang University, Beijing 100191, China

[2]Big Data Center of Student Affairs Department, Beihang University, Beijing 100191, China

[3]School of Instrumentation Science and Opto-electronic Engineering, Beihang University, Beijing 100191, China

*Corresponding author.

Keywords: Online behavior, Higher education, Big data.

Abstract. Internet has become almost universal among university students, which has greatly impacted their daily life and learning performance. In this paper, by studying the data regarding internet use and academic achievement during four years, online behavior of undergraduates from university B in terms of students' type and online period was characterized. Then, a simplified principal component analysis method for interval data was employed to analyze the correlation between online behavior and academic record. Results indicated the dependence of students on the internet increases with grade growth. Moreover, excessive internet access may bring negative effects on learning performance.

大数据视角下B高校本科生校园网络行为分析

刘科生[1,2]，王梦醒[2,3]，倪义坤[1,2,*]

[1]北京航空航天大学经济管理学院，北京，中国

[2]北京航空航天大学学生大数据中心，北京，中国

[3]北京航空航天大学仪器科学与光电工程学院，北京，中国

[*]通讯作者

关键词：网络行为；高等教育；大数据

摘要：互联网与当前大学生的学习、生活息息相关。本文以B高校2018届本科毕业生大学四年校园网络数据和成绩数据为例，考察了不同类型学生和学生在不同时间段的网络行为特点，并通过区间主成分分析方法进一步研究了校园网络行为和学习成绩的相关性。主要发现随着年级的增长，学生对网络的使用和依赖度逐渐增加，过度使用网络（高上网流量和时长）总是会对学习成绩产生严重影响。

1．引言

现代信息技术在高等教育领域的广泛应用，为我国高等教育转型发展带来了新机遇，同时也对高等教育理念、教育模式产生了深刻影响。2018年《教育部关于加快建设高水平本科教育全面提高人才培养能力的意见》中指出：要大力推动互联网、大数据、人工智能、虚拟现实等现代技术在教学和管理中的应用……以现代信息技术推动高等教育质量提升的"变轨超车"[1]。在此背景下，积极探索互联网、大数据技术在高等教育中应用的新途径、新方法，对努力提升高等教育工作科学化、精准化，具有重要的理论价值和实践意义。

目前，“95后”甚至“00后”已经成为大学生主体，互联网的浸润不仅使他们变得更加开放、活跃、追求自由、崇尚个性，也使他们的生活、学习、社交等活动与互联网密不可分。对学生网络行为数据的分析，有助于把握新时代学生的思想、行为特征，客观的了解大学生网络行为的特征和规律，剖析网络行为与大学生成长成才的关系，从而能够进一步提升对学生教育、指导的针对性和科学性，有利于提升高等教育的科学化水平。

近年来，教育与大数据的交叉融合在国际上获得了广泛关注。2008年，国际教育数据挖掘委员会指出：教育数据挖掘是一个将来自教育系统的原始数据转换为有用信息的过程，这些有用信息可以为教师、学生、家长、教育研究人员及教育软件开发人员所用[2]。然而，数据挖掘在我国教育领域的应用仍然有限，例如，国内对大学生网络行为的研究主要基于问卷调查的方式，虽然取得了一系列研究进展，但在数据的真实性、完整性及周期性等方面存在局限，因此所得结论在一定程度上也可能存在偏差[3]。

本文以北京地区B高校2018届本科毕业生为研究对象，通过收集并整理其大学四年期间与校园网络行为相关的全部数据分析统计规律，在此基础上深入挖掘了校园网络行为与学习成绩的相关性，并结合实际情况给出思考与建议。

2．样本选取与数据来源

2.1 样本选取

本文选取的样本为北京地区B高校2018届3815名本科毕业生，包括3014名男生与801名女生，平均年龄为22岁。已有统计数据表明，约90%的B高校学生自小学甚至小学以前就已经开始接触互联网，互联网已经融入他们生活、学习的点点滴滴。此外，作为一所地处首都的一流理工科学府，B高校不仅有能力将学生在大学期间产生的数据完整保存下来，而且能够为学生提供更完善的互联网服务。因此，B高校学生在大学生网络行为研究方面十分具有代表性。

2.2 数据来源

本文使用了B高校2018届本科毕业生大学八个学期的数据库，涵盖该校全体2018届本科生大学四年全部校园网络行为与学习成绩的数据。其中，校园网络行为数据由网络管理系统中提取的23843813条记录构成，包括流量、在线频率、上线时间和下线时间等；学习成绩数据指加权平均分，根据学生的学习成绩与学分计算，由教务管理系统提供。需要指出的是，在本研究开始前，我们已经删除了原始数据中可能涉及隐私的相关信息。

3．研究方法

3.1 数据清理与相关性分析

本文首先对B高校2018届本科毕业生在校四年的校园网数据进行了清理和统计，初步统计的变量为：上网流量、上行流量、下行流量上线频率和上线时长五个变量。

在统计分析的基础上，将以上5个变量与学生加权平均分进行相关性分析，采用皮尔森相关系数（Pearson correlation coefficient）和斯皮尔曼等级相关系数（Spearman's rank correlation coefficient）两种相关系数[4]：皮尔森相关系数是一种线性相关系数，用来反映两个变量线性相关程度。相关系数描述的是两个变量间线性相关强弱的程度。其相关系数的绝对值越大表明相关性越强；斯皮尔曼等级相关系数用单调函数来描述两个变量之间的相关性，当其中一个变量可以表示为另一个变量的很好的单调函数时（即两个变量的变化趋势相同），两个变量之间的相关系数可以达到+1或-1。

3.2 主成分分析

本文中涉及5个变量的分析研究，为了进一步得到分析各变量与学习成绩之间的关系，采用主成分分析（Principal component analysis，PCA）研究各变量与学习成绩的关系。主成分分析（Principal component analysis，PCA）是数据挖掘最常用的“降维”方法之一。本文使用了基于简化的区间数据主成分分析（Simplified principal component analysis，SPCA），它具备信息有效利用和简单操作的优点。SPCA的简要计算步骤如下[5, 6]：

设一个由N个样本点和 p 个变量构成的数据表 $X_{N\times p}=(x_{ij})_{N\times p}$，

第1步：求解$X_{N\times p}$的相关系数矩阵 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$和特征向量$u_1,u_2,\cdots,u_p$，其中$u_h=(u_{h1},u_{h2},\cdots,u_{hp})'$和$u_1,u_2,\cdots,u_p$是正交的。

第2步：根据累计贡献率确定保留主成分的个数m，累计贡献率的确定方式如下：

$$Q_m=\frac{\sum_{h=1}^{m}\lambda_h}{\sum_{j=1}^{p}\lambda_j} \tag{1}$$

第3步：求区间主成分$F_1,F_2,\cdots,F_m$。若记$F_h(i)=[\underline{f}_{ih},\bar{f}_{ih}]$是一个区间数据，它是第$i$个样本点在第$h$个主成分上的取值，根据Moore提出的区间数据线性组合算法，有：

$$\underline{f}_{ih}=\sum_{j=1}^{p}u_{hj}\left[\tau\underline{x}_{ij}+(1-\tau)\bar{x}_{ij}\right] \tag{2}$$

$$\bar{f}_{ih}=\sum_{j=1}^{p}u_{hj}\left[(1-\tau)\underline{x}_{ij}+\tau\bar{x}_{ij}\right] \tag{3}$$

其中$\tau=\begin{cases}0,u_{ij}\leq 0\\1,u_{ij}>0\end{cases}$

由此可得主成分为：$F_h=(F_h(1),F_h(2),\cdots,F_h(n))',h=1,2,\cdots,m.$

4．结果与讨论

4.1 “95”后大学生网络行为特点

4.1.1 互联网使用程度明显增高

通过对2018届本科毕业生的校园网络行为数据进行统计，可以发现本届毕业生人均上网流量为2.28 GB/天，为2017届（1.08 GB/天）的2倍多，大学生每天的上网流量呈现出逐年增高的趋势。这在一定程度上说明，大学生的信息化程度日益提升，互联网在大学生的学习与生活中扮演着原来越重要的角色。

4.1.2 男生上网流量明显高于女生

2018届本科毕业生中，男生的人均上网流量为2.45 GB/天，女生为1.43 GB/天。究其原因，可能与男生和女生网络行为的特点有关。男生更偏向网络游戏，下载视频、软件和资料等；女生相对安静感性，更偏向于使用网络社交、购物以及资讯浏览等。

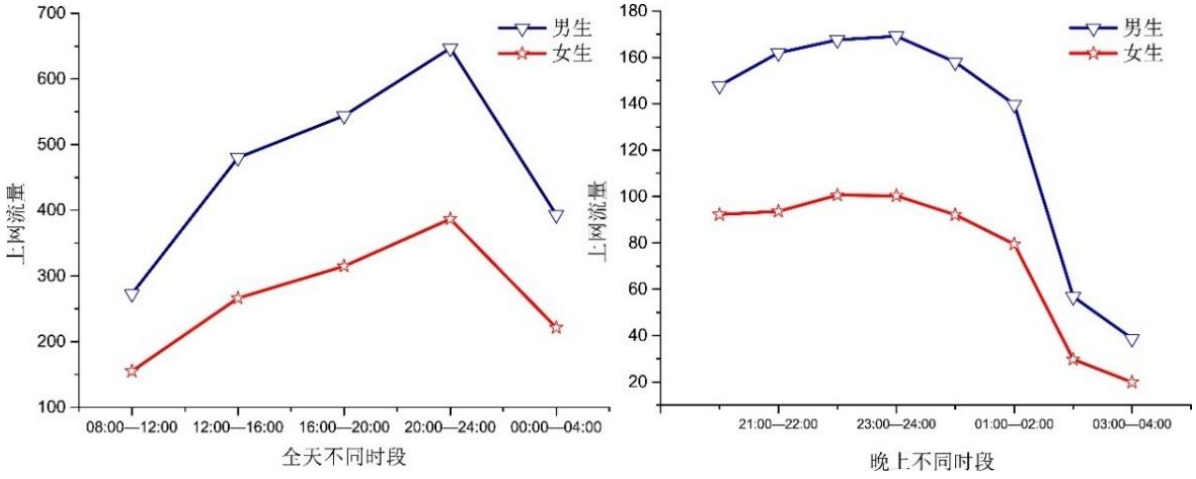

图1　男生、女生不同时段上网流量变化

4.1.3 工科学生上网流量高

通过分析不同学院学生的校园网络行为，可以发现工科学生每天的上网流量最高，其次是理科，文科学生每天的上网流量相对较少。

4.1.4 喜欢在特定时间上网

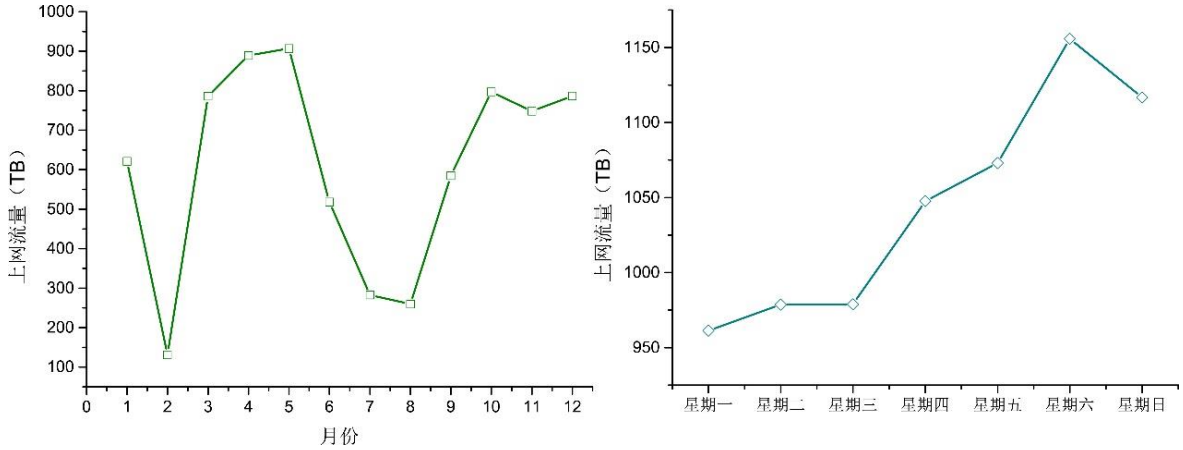

图2　2018届本科生不同月份及每周上网流量变化

每年中，春季学期的上网流量高于秋季学期。每周内，周末的上网流量高于平日，周六上网流量最高。时间段分布方面，男女生人均上网流量每天的变化趋势和晚上每小时的变化趋势基本一致，并均在22:00-24:00最高。

4.2　网络行为与学习成绩的相关性

表1　各变量与学习成绩相关系数

变量	相关系数	皮尔森检验 Pearson's test	斯皮尔曼检验 Spearman's test
总流量	相关系数	-.160	-.169
	显著性	.000	.000
上行流量	相关系数	-.164	-.168
	显著性	.000	.000
下行流量	相关系数	-.097	-.148
	显著性	.000	.000
上线频率	相关系数	.044	.062
	显著性	.000	.000
上线时长	相关系数	-.009	-.037
	显著性	.159	.100

尽管本文中采用两种相关性分析方法，但是两种方法得到了一致的结果：除上线频率外，其他四个在线行为指标均与加权平均分呈负相关，其中上网流量与加权平均分的相关系数绝对值明显大于其他指标。从相关性分析中可以看到，上网流量是影响加权平均分的主要因素，因此过度使用互联网确实会对学生的学习成绩产生一定影响。

接下来本文利用SPSS (Statistical Product and Service Solutions, IBM, USA) 区间主成分分析，分析对象为上网流量、上行流量、下行流量上线频率和上线时长5个变量，在产生的主成分中，前两位主成分的累积贡献率为84.14%。表2反映了前两个主成分和5个变量的相关性。

表2　各主成分与变量相关性

	第一主成分	第二主成分
总流量	0.963	0.243
上行流量	0.834	0.355
下行流量	0.858	-0.043
上线频率	0.023	0.912
在线时长	0.303	0.859

从表2 中可以看出，第一主成分与上线频率和在线时长相关性较弱，与总流量、上行流量和下行流量相关性较强，因此第一主成分是反映学生的上网流量整体情况的指标。与此相反，第二主成分则与上线频率和在线时长相关性较强，与上总流量等其余变量相关性较弱，因此第二主成分可作为上线频率和时长的总体指标，可以反映学生对网络的依赖度。

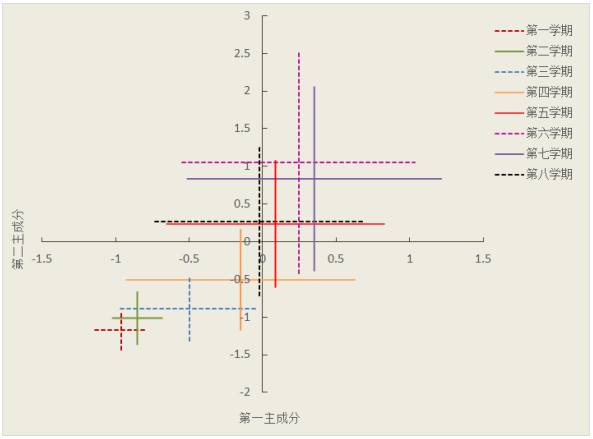

图3　各学期上网情况的主平面图

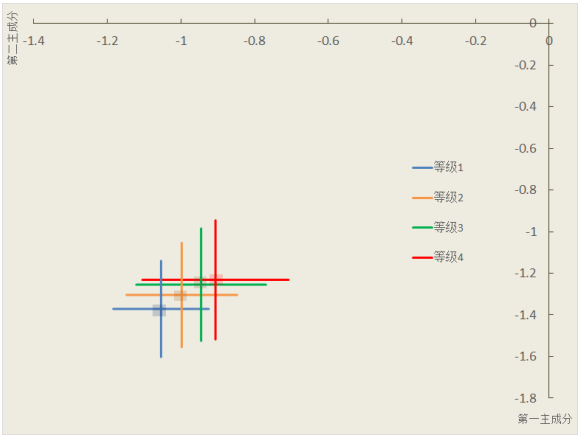

图4　不同成绩段的主平面图

将学生各学期的上网数据投影到第一主成分和第二主成分组成的主平面上，可得到如图3所示的主平面图。

图3表示2018届本科毕业生大学四年各学期的上网情况的区间主成分分析结果，其中第一主成分为纵轴，第二主成分为横轴。本文采用第一个四分位数和第三个四分位数作为区间的下限和上限。图中的“十”字代表了学生在不同学期的网络使用情况，其中每一个“十”字的交叉点表示相应学期学生校园网络使用情况的平均水平，而“十”字的长短的表示了各学期学生网络使用情况在第一主成分和第二主成分上的离散范围。因此，通过观察各个“十”字的位置，可以清晰的了解学生在不同学期的上网情况。

从图3中可以看到，八个学期可分为四部分，这恰好对应于四个学年，代表第一学期和第二学期的“十”字在该图的两个维度中显示最低分数，说明大一学年学生的校园网络使用程度较低。随着年级的提升，两个成分的分数逐渐增加，代表学生最后两个学期的“十”字得分最高。由此可以看到，随着年级的提高，上网流量逐渐增高，对网络使用和依赖程度也逐渐增高。

本文将学生的成绩划分为四个等级，等级1表示加权平均分最高，等级4表示加权平均分最低。将学生的成绩投影到第一主成分和第二主成分组成的主平面上，可得到如图4所示的主平面图。

在图4中可以看到，等级1代表的具加权平均分最高的学生群体，在两个维度中显示了最低分数。随着课程成绩的下降，两个维度的分数逐渐增加。因此可以说明，这意味着使用较少在线流量并在互联网上花费较少时间的学生更有可能获得好成绩。

5．结论

现在网络与大学生学习、生活密切相关，各高校都尽可能满足学生的需求，帮助学生拓宽视野，拓展知识渠道。B高校2018届本科毕业生大学四年中各学年的在线流量平均分别为0.43GB，1.43GB，1.79GB和2.01GB，作为进入大学的新生，大学生基础课程学习压力较大，在互联网上花费的时间相对较少。随着年级的提高，由于专业课的开展和个人发展的需求，他们会花费更多的精力从互联网获取信息和资源，但同时也有部分学生花费很多精力在网络游戏、网络小说等活动中，这些原因会导致随着学期的增长，网络的使用度和依赖程度越来越高。

大学生几乎不会受到类似于高中时期的约束和限制，在紧张的学习环境中，一些学生一旦沉迷于网络，往往很容易失去学习兴趣，从而形成恶性循环。本文中可以发现在线流量和

在线时长是影响所有指标中课程成绩的主要因素。从等级1到等级4，每日平均在线流量分别为0.26GB，0.31GB，0.39GB，0.45GB，平均每日在线时长分别为4.93h，5.48h，5.96h，6.15h。学生在网络上花费越多，他们对学习的关注就越少，从而导致学习成绩不佳。从在线频率的角度来看，不同课程成绩的学生之间没有显着差异。具有良好学习成绩的学生也经常使用互联网，但他们认为互联网是一种更好的方式，可以帮助他们获得更多的学习知识资源，而这些知识不需要太多的流量和时间，每次登录网络时，他们花费的时间和流量相对较少。因此，合理的使用互联网有助于帮助学生提高学习成绩，过度使用网络总是会对学习成绩产生严重影响。

References

[1] ‘Opinions of the Ministry of Education on Accelerating the Construction of High-level Undergraduate Education and Improving the Ability of Talent Cultivation’, 2018.

[2] Merceron A., Educational data mining/learning analytics: Methods, tasks and current trends[J]. 2015.

[3] Ting Li, Gang-shan Fu, An overall view of the educational data mining domain[J]. Modern Educational Technology, 2010. 20(10): 21-25.

[4] Xiao-yan Wang, Mei-zhou Li, The relationship of rank correlation coefficient and spearman rank correlation coefficient [J]. Journal of Guangdong Industry Technical College, 2006. 5(4): 26-27.

[5] Moore R.E., Interval analysis[J]. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

[6] Yin Zhang, Yan Wang, Hui-wen Wang, Evaluating of academic journals in management of key academic journal fund: An application of simplified principal component analysis based on interval data[J]. Journal of Management Sciences in China, 2010. 13(7): 88-94.