大数据时代高校学生阅读行为分析研究

Analysis and Research on College Students' Reading Behavior in the Era of Big Data

王弘江[1] 韩明洁[2] 刘尚懿[2*]

Hongjiang Wang[1] Mingjie Han[2] Shangyi Liu[2*]

1. 辽宁科技大学材料与冶金学院 中国·辽宁 鞍山 114051
2. 辽宁科技大学计算机与软件工程学院 中国·辽宁 鞍山 114051

1.School of Materials and Metallurgy, Liaoning University of Science and Technology, Anshan, Liaoning, 114051, China
2.School of Computer and Software Engineering, Liaoning University of Science and Technology, Anshan, Liaoning, 114051, China

摘 要：传统学生行为分析中的采样困难、样本覆盖面小。在大数据时代背景下对高校学生行为进行分析，采集高校学生的全部日常行为数据。在此基础上，利用大数据分析算法对采集到的学生行为数据进行分析，发现数据背后的规律，进而分析出学生的行为动机，并根据分析到的学生行为，建立对策数据库。实验对比结果表明，此次设计的高校学生行为分析方法比传统的方法准确性高，具有一定的实际应用意义。

Abstract: In traditional student behavior analysis, sampling is difficult and sample coverage is small. In the context of the era of big data, to analyze the behavior of college students, all the daily behavior data of college students is collected. On this basis, the big data analysis algorithm is used to analyze the collected student behavior data, discover the rules behind the data, and then analyze the behavior motivation of students, and establish the countermeasure database according to the analyzed student behavior. The experimental results show that the design of college student behavior analysis method is more accurate than the traditional method, has a certain practical significance.

关键词：大数据时代；学生行为；阅读

Keywords: big data era; student behavior; reading

DOI: 10.12346/sde.v3i10.4544

1 引言

在大数据时代，学习大数据的知识，掌握大数据分析的方法，是大学生必备的技能[1]。回首过去，从幼儿园到大学，一直都在做着一件事，那就是“读书”，读书是提高个人素质的重要途径，是学习过程中重要的环节。大学阶段是读书观形成并逐步稳定的关键时期，对于大学生的心理健康、人格养成和社会适应等都具有重要的意义。为了了解大学生的读书现状及其存在的问题，唤起大学生的读书意识，对辽宁科技大学图书馆的借阅记录的大数据进行了分析研究，发现了许多有趣的现象。在对这些现象进行简单的分析后，发现了现在大学生的阅读兴趣和行为模式，为学校对学生进行针对性的引导与教育、制定相关决策提供了客观依据。

本次研究使用的硬件平台为 HP Proliant 服务器，CPU 至强 5620，内存 32GB，软件平台为 Ubuntu 18 系统，主要编程语言为 Python。使用 Python 生态系统中著名的数据分析软件包 Pandas 与数据可视化软件包 Matplotlib[2]。

2 数据与方法

2.1 数据的收集与整理

为了对借阅数据进行多角度的分析，需要其他相关的数据，如需要图书分类信息、学校院系编码信息等。由于这些数据都比较小，所以直接通过网络资源进行了人工的收集与整理。《中国图书馆分类法》是中华人民共和国成立后编制出版的一部具有代表性的大型综合性分类法，是当今中国图

【作者简介】王弘江（1999-），男，中国四川泸州人，在读本科，从事材料加工及控制、科学计算研究。
【通讯作者】刘尚懿（1976-），中国辽宁鞍山人，硕士，讲师，从事图像处理、深度学习研究。

书馆使用最广泛的分类法体系，简称《中图法》。《中图法》已普遍应用于全国各类型的图书馆，中国主要大型书目、检索刊物、机读数据库以及《中国国家标准书号》等都著录《中图法》分类号。部分分类结果如图 1 所示。

2.2 数据读取与清洗

数据读取与清洗是大数据分析过程的第一步。由于现实世界的数据常常是不完全的、有噪声的、不一致的，所以需要对这些数据进行修补或移除以提高数据质量，数据清洗对随后的数据分析非常重要，因为它能提高数据分析的准确性 [3]。本次研究获得了从 2016 年 1 月 1 日至 2020 年 5 月 1 日的全部借阅记录数据，共 80 多万（817742）条记录。保存在 cvs 类型文件中。将所有列名称转为小写字母形式，将所有数据转换为字符串形式。出于保护隐私的目的，首先删除“学生姓名、书籍作者、出版商”字段，再删除“国际标准书号、续借日期”等冗余字段。通过学号判断学生的入学日期，清洗后的部分数据如表 1 所示。

2.3 数据分析方法

按照一般数据分析原则，首先对数据进行现状的分析和对比。我们通过对比过去发生的事情去了解现在所处的情况。然后再进行原因分析 [4]。我们对每次分析得出的现象进行原因分析，分析出这一现象的起因和影响要素。

采用的分析方法：

①主要采用分类的方法，将数据中的某组数据按照共同特点进行分类，通过分类模型获得某项数据的类别特征。

②还采用了聚类的方法，将数据中按照相似性和差异性分成不同的类别，通过聚类模型得到相似的类别特征 [5]。

3 数据分析及结果

3.1 宏观分析，整体认识数据

为了全面细致地对高校学生阅读情况进行大数据分析，首先统计出男女学生人数，再根据性别进行分组统计，得到男女学生的借阅书籍的总数量，如图 2 所示 [6]。

①在 2016 年 1 月 1 日—2020 年 5 月 1 日期间，该校共有 25938 名学生（含 2016 年毕业生）进行了借阅，男女生比例为 1.6 ： 1。

②在大学四年中每个男生从图书馆共借阅 19.2 本书籍，每个女生共借阅 22.3 本书籍，男女生比例为 1 ： 1.16。

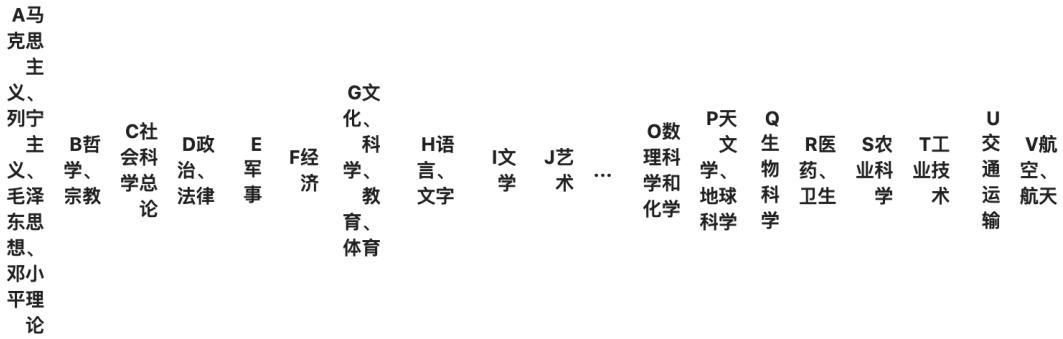

图 1 中图分类法

表 1 清洗后数据

	cert_id	dept	sex	m_title	m_call_no	lend_date	ret_date
0	120173605073	财务171	F	创新设计:TRIZ：发明问题解决理论	TB472/T2201～3	2018-11-1310:13:41	2018-11-2715:06:05
1	120174502004	金融工程17	F	美国货币史:1867-1960:1867-1960	F827.12/160001/1-2	2018-11-1818:31:49	2018-11-2411:57:05
2	120183803046	信息18	F	DK·牛津英汉双解大词典	H316/2Y572/1～2	2018-10-1920:13:28	2018-11-1818:32:38
3	120164801007	无机2016-1	M	明朝那些事儿.第六部,帝国,山雨欲来	K248.09/D221:6/3～3	2018-11-1818:34:58	2018-12-0110:40:07
4	120164801007	无机2016-1	M	明朝那些事儿.第五部,内阁不相信眼泪.典藏本	K248.09/D221:5/3～3	2018-11-1818:34:59	2018-12-0110:40:08

3.2 按年级进行统计分析

对借阅数据按照入校的年份进行分组统计，得到每年入学学生的全部借阅情况，对比不同入学年份学生的借阅情况，如图 3 所示。

通过分析得到图 3，从中可以看出：

① 2019 年入学的新生借阅图书的数量猛增，达到 4742 人次，接近当年入校学生（4863）的 97.5%，差不多每人借阅过一次书籍。这是因为图书馆针对 2019 年入学的新生开设了一门《信息检索》的公开课导致的。

②从 2 年级学生的借阅数据可以看到，经过 1 年的大学生活，还有接近 45% 的学生没有在图书馆借阅过至少 1 本书籍。说明学生对图书馆中资源的利用率不高。

③学生借阅图书的累计数量逐渐增加，平均一年不足 10 本（8.25 本），低于大学学生平均每年读 18 本电子书的结论，说明现在的学生更愿意阅读电子书籍而不是纸质书籍。

3.3 按照学院分组进行分析

按照各个学院分别进行分析，得到不同学院学生借阅情况，分析过程如下：

①根据学生‘cert_id’字段中的学院编码规则，提取出学生所在学院的编码。

②按照编码不同进行分组统计，得到数据。

③根据学院编码表对数据进行分组，统计每个学院学生的总数量和学生数量，求出每个学院的平均借阅数量，以柱状图方式展示，如图 4 所示。

通过上述分析可以看到：

①国际金融与银行学院的平均借阅数最高，应用技术学院的平均借阅数最低。

②理工科学院（化工学院、机械学院、材料学院、电信学院等）的平均借阅数基本持平。

③特色办学的学院（软件学院、艺术学院）的借阅数差不多。

3.4 统计学生阅读书籍种类

根据学生‘cert_id’字段中的学院编码规则，提取出学生所在学院的编码；按照编码不同进行分组统计，得到数据。根据学院编码表对数据进行分组，统计每个学院学生的总数量和学生数量，求出每个学院的平均借阅数量，以柱状图方式展示，得到图 5。

通过分析，我们可以得到如下结论：

①学生最喜欢阅读的书籍的第一名是中国文学类，主要是各种小说。

②学生最喜欢阅读的书籍的第二名是计算机类，主要是由于计算机二级考试带来的影响；同时，由于各个专业需要学习很多计算机方面的知识和技术；再者，因为学校学生中男生比女生多，男生更喜欢计算机类书籍。这些共同因素使得计算机类的书籍广受欢迎。

③学生最喜欢阅读的书籍的第三名是外国语类，主要是各种英语类书籍，因为大学英语四六级考试的影响很大。

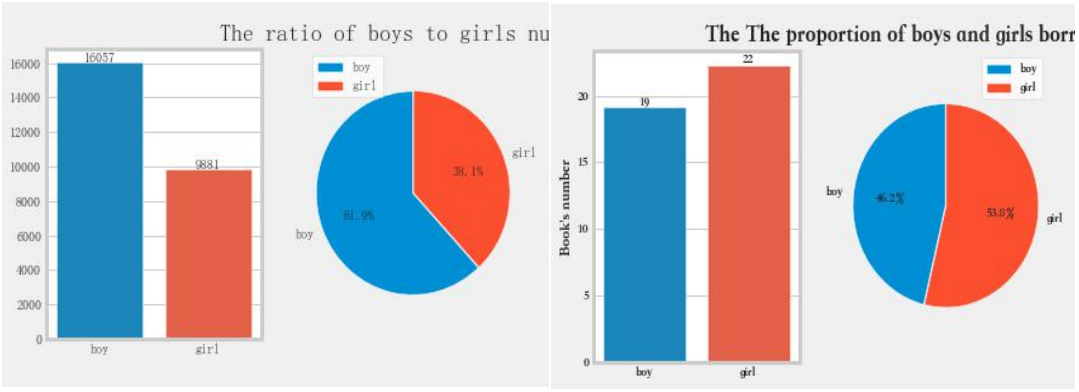

图 2 男女学生借阅情况

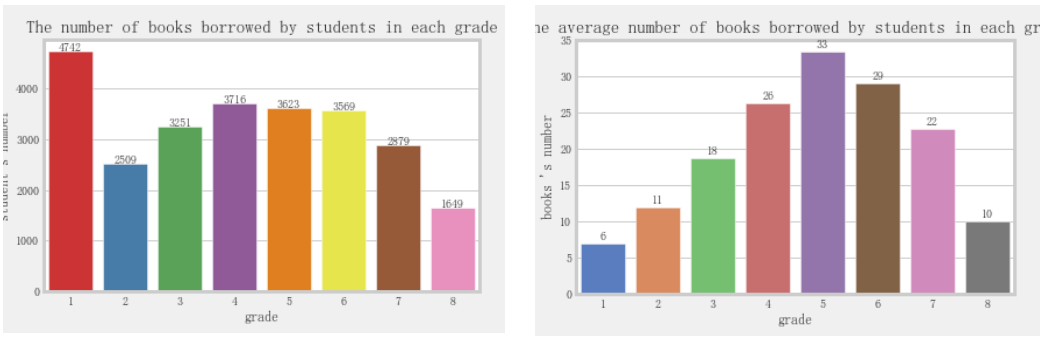

图 3 不同年级借阅情况

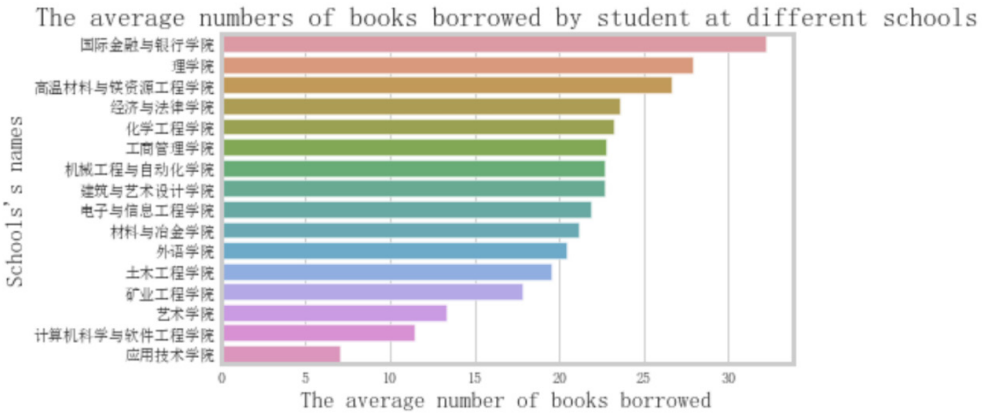

图 4 不同学院借阅情况

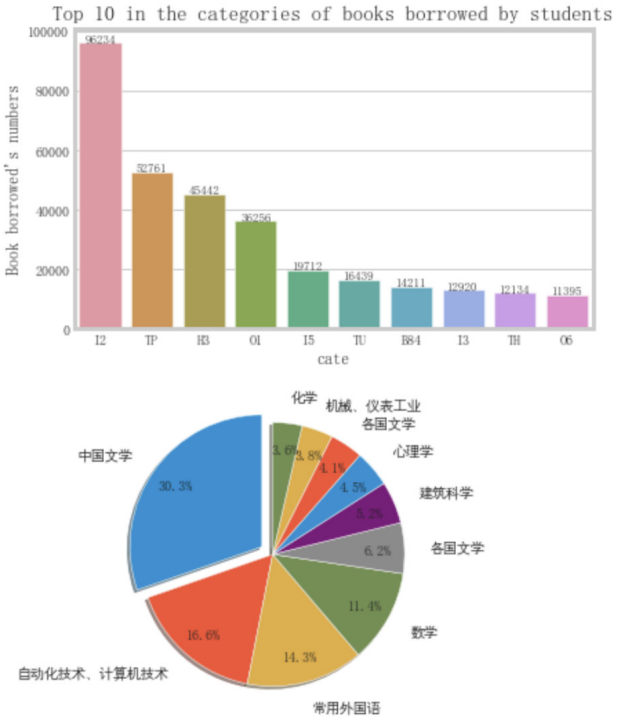

图 5 不同种类图书借阅情况

4 总结与讨论

在本课题中，我们对数据几个方面进行的统计与分析。分别从数据的不同角度进行了数据分组与数据聚类，得到了许多有价值的现象和结论，从中了解了目前大学生的整体阅读行为及其背后的原因。为学校更好地引导与教育学生，更细致地为学生提供服务提供了客观依据。是一次非常有意义的尝试。今后应定期进行这样的大数据分析。

4.1 不足之处

虽然本次研究得到了比较丰富的成果，但是还是有所欠缺。主要原因是分析的数据来源不够丰富，主要数据只有图书馆的借阅数据，还缺乏教务处掌握的历年学生成绩数据，后勤管理处掌握的学生生活消费数据等。如果通过这些数据与图书馆数据进行联合分析，我们相信一定会得到更多有价值，能说明更多问题的结果，这对更好地了解学校学生的状况，提前预测学生的行为，实现更精细化的服务一定是大有帮助的。

4.2 后续工作

在本课题的后续阶段，我们准备扩大数据的来源，通过采集学生学习成绩数据、学生生活消费数据、学生社交信息数据等多维度数据，分析这些信息之间的相关度，全方位了解学生学习、生活、思想等多方面的状态。

参考文献

[1] [美]托马斯•埃尔.大数据导论[M].北京:机械工业出版社,2017.

[2] 余本国.基于Python的大数据分析基础及实战[M].北京:水利水电出版社,2018.

[3] 黄源,蒋文豪,徐受蓉.大数据分析:Python爬虫、数据清洗和数据可视化[M].北京:清华大学出版社,2020.

[4] 刘凯悦.大数据综述[J].计算机科学与应用,2018,8(10):1503-1509.

[5] 张锋军.大数据技术研究综述[J].通信技术,2014,47(11):1240-1248.

[6] 任磊,杜一,马帅,等.大数据可视分析综述[J].软件学报,2014,25(9):1909-1936.