



重庆交通大学  
CHONGQING JIAOTONG UNIVERSITY

# 本科毕业论文（设计）

题目： 基于聚类算法的校园网用户行为分析系统设计与实现

学 院： \_\_\_\_\_

专 业： \_\_\_\_\_

学 生 姓 名： \_\_\_\_\_

学 号： \_\_\_\_\_

指 导 教 师： \_\_\_\_\_

评 阅 教 师： \_\_\_\_\_

完 成 时 间： \_\_\_\_\_

重庆交通大学

CHONGQING JIAOTONG UNIVERSITY

## 本科毕业论文（设计）原创性声明

本人郑重声明：所提交的毕业论文（设计），是本人在导师指导下，独立进行研究工作所取得的成果。除文中已注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文研究做出过重要贡献的个人和集体，均已在文中以明确方式标明。

本人完全意识到本声明的法律后果由本人承担。

作者签名（亲笔）：

年 月 日

---

## 本科毕业论文（设计）版权使用授权书

本毕业论文（设计）作者完全了解学校有关保留、使用学位论文的规定，本科生在校攻读期间毕业论文（设计）工作的知识产权单位属重庆交通大学，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权重庆交通大学可以将毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编毕业设计（论文）。

作者签名（亲笔）：

年 月 日

导师签名（亲笔）：

年 月 日

## 摘 要

随着互联网的发展,校园网也在各大高校得到了广泛的使用,越来越多的人享受到了网络所带来的便利。但同时,网络用户数量的不断增长以及用网频率的增加也为校园网络管理带来了许多新的问题。如何对校园网络用户所产生的海量数据进行有效的数据挖掘,对高校的网络与教学管理都有着重大的意义。用户网络日志往往伴随着高维度,且部分属性间具有相关性的特点。本文旨在根据校园网学生用户的网络日志数据来开发一个基于聚类算法的校园网用户行为分析系统,该系统以实际校园网络数据为实验数据,对用户的网络依赖程度、用户访问网站类型的偏好、总体的网络使用情况进行了分析,从而实现对校园网络以及学生更好的管理。论文主要工作内容如下:

针对校园用户日志数据的多维性,且部分属性具有相关性的特点,提出了有效的预处理办法,对不同属性进行了有效的融合分析。

概述了几种主流的聚类技术,并特别针对 K-means++、DBSCAN 和自编码器这三种聚类方法,探讨了它们在分析用户行为模式时的聚类效果。针对聚类结果进行了详细的分析。

完成了可视化系统的设计与实现,为校园网管理者提供了有效的网络数据分析工具,帮助其更好地维护和管理网络。同时,借助此系统辅导员能够更能清晰明了地掌握学生对网络的使用情况,从而能够做出更迅速、准确的决策和管理。

**关键词:** 校园网; 用户行为分析; 数据挖掘; 聚类算法

## **Design and Implementation of Campus Network User Behavior Analysis System Based on Clustering Algorithms**

### **Abstract**

With the development of the Internet, the campus network has also been widely used in various colleges and universities, and more and more people enjoy the convenience brought by the network. But at the same time, the increasing number of network users and the increase of the frequency of network use has also brought many new problems for the campus network management. How to effectively mine the data of the massive data generated by the campus network users is of great significance to the network and teaching management of colleges and universities. User network logs are often accompanied by high dimensions, and some attributes have the correlation characteristics. This paper aims to the campus network students network log data based on algorithm to develop a clustering network user behavior analysis system, the system to the actual campus network data, the user network dependence, the user access site type preference, the overall network usage is analyzed, so as to realize the better management of campus network and students. The main work contents of the paper are as follows:

In view of the multi-dimensional nature of campus user log data and the relevance of some attributes, we propose an effective preprocessing method and make an effective fusion analysis of different attributes.

Several mainstream clustering techniques are outlined, and the three clustering methods of K-means + +, DBSCAN and autoencoders are specifically targeted to explore their clustering effects when analyzing user behavior patterns. A detailed analysis was conducted for the clustering results.

The design and implementation of the visualization system were completed, providing effective network data analysis tools for the campus network managers to help them better maintain and manage the network. At the same time, with the help of this system, the counselors can more clearly grasp the students' use of the network, so as to make more rapid and accurate decisions and management.

**Keywords:** Campus Network; User Behavior Analysis; Data Mining; Clustering Algorithms

## 目 录

摘 要 .....	II
Abstract .....	III
第 1 章 绪论 .....	1
1.1 选题背景及意义 .....	1
1.2 国内外研究现状 .....	1
1.2.1 国外研究现状 .....	1
1.2.2 国内现状 .....	2
1.3 研究内容 .....	3
1.4 论文结构安排 .....	3
1.5 本章小结 .....	4
第 2 章 相关理论与技术 .....	5
2.1 预处理方法 .....	5
2.1.1 数据清洗 .....	5
2.1.2 数据变换 .....	5
2.1.3 数据标准化 .....	5
2.1.4 数据加权归一化 .....	7
2.2 聚类算法介绍 .....	7
2.2.1 基于划分的聚类算法 .....	8
2.2.2 基于密度的聚类算法 .....	9
2.2.3 基于自编码器的聚类算法 .....	9
2.3 系统开发相应技术 .....	11
2.3.1 前端展示层: Vue.js .....	11
2.3.2 后端服务层: Java Spring Boot .....	12
2.3.2 脚本自动化: Script .....	12
2.4 本章小结 .....	12
第 3 章 预处理及聚类算法实现 .....	13
3.1 校园网用户行为分析背景 .....	13
3.1.1 数据获取 .....	13
3.1.2 数据特点 .....	14
3.2 数据预处理 .....	14
3.2.1 日志数据清洗 .....	14

3.2.2	日志数据变换 .....	14
3.2.3	日志数据标准化 .....	17
3.2.4	日志数据归一化 .....	17
3.3	聚类效果对比及结果分析 .....	17
3.3.1	自编码器训练 .....	18
3.3.2	聚类效果对比 .....	18
3.3.3	聚类结果分析 .....	19
3.3	本章小结 .....	25
第 4 章	系统需求分析与设计 .....	26
4.1	系统需求分析 .....	26
4.1.1	功能性需求分析 .....	26
4.1.2	非功能性需求分析 .....	26
4.2	系统设计 .....	27
4.2.1	系统开发环境 .....	27
4.2.2	系统功能设计 .....	27
4.2.3	数据库设计 .....	28
4.3	本章小结 .....	31
第 5 章	系统实现 .....	32
5.1	登录模块 .....	32
5.2	注册模块 .....	34
5.3	协议信息展示模块 .....	35
5.4	用户网络依赖程度展示模块 .....	38
5.5	用户访问偏好展示模块 .....	39
5.6	基本信息展示模块 .....	40
5.7	本章小结 .....	40
第 6 章	系统测试 .....	41
6.1	系统测试的目的 .....	41
6.2	系统测试的方法 .....	41
6.3	系统测试实现 .....	41
6.4	本章小结 .....	50
第 7 章	总结与展望 .....	51
致    谢	.....	53

参 考 文 献 .....	54
---------------	----

## 第 1 章 绪论

### 1.1 选题背景及意义

根据中国互联网络信息中心（CNNIC）发布的第 52 次《中国互联网络发展状况统计报告》（2023 年 8 月 23 日），截至 2023 年 6 月，中国网民规模达到 10.79 亿，互联网普及率达 76.4%。人均每周上网时长为 29.1 个小时<sup>[1]</sup>。这显示出人们与互联网之间的关系愈发紧密。互联网为人们提供了各种活动和资源的途径。通过互联网，人们可以在社交媒体上与朋友和家人保持联系，使用电子邮件和即时通信工具进行沟通，获取新闻和实时信息，参与在线教育和学习，进行购物和支付，享受娱乐和游戏，并搜索和获取各种知识和资源。互联网的广泛应用极大地改变了人们的生活方式和工作方式，使得人们可以方便地进行交流、获取信息和进行各种活动。校园网络作为互联网的关键组成部分，在校园教育领域发挥着举足轻重的作用。但是，随着校园网络用户数量的激增，也为网络的使用和管理带来一系列挑战。一些学生沉迷于网络游戏、色情视频等，导致学习成绩下降，或在赌博网站上输掉大量财物。因此，近年来，人们越来越关注校园网用户行为<sup>[2]</sup>。

通过分析和提炼大量学生的网络行为数据，我们能够发现学生群体的上网习惯以及校园网络的整体运作模式。这不仅能够为校园网络的构建、维护和优化提供决策依据，还能为教育教学管理提供数据支持<sup>[3]</sup>。目前，用户行为分析常使用统计分析、聚类分析、关联规则挖掘、机器学习和用户反馈调查等方法，以研究和理解用户在特定环境下的行为模式和趋势<sup>[4]</sup>。其中，聚类算法是应用较广泛的方法之一。因此，设计一套基于聚类算法的校园网用户行为分析系统对于校园网用户日志数据来说是非常有意义的。

### 1.2 国内外研究现状

#### 1.2.1 国外研究现状

国外对于校园用户行为分析的研究相对于国内起步更早，其中 Xhafa F 等人在虚拟校园这一教育场景中，通过应用 abi 双聚类算法，对成千上万学生和导师在在线学习平台上产生的庞大、复杂且异构的活动日志数据集进行了深入分析，成功提取了用户的



导航模式、执行的活动类型以及相关的时间参数等关键信息，为虚拟校园的设计者和开发人员提供了关键的数据支持，帮助他们更好地理解用户行为，从而推动了在线教育平台的持续改进和发展。Singh M<sup>[5]</sup>提出了一种集成混合机器学习模型，该模型融合了多状态长短期记忆网络（MSLSTM）和卷积神经网络（CNN）技术，用于对用户行为操作序列进行深入的时间序列异常检测，以识别和防范组织内部的恶意行为，并通过在公开的内部威胁数据集上的实验，展示了该模型在训练集上达到 0.9042、测试集上达到 0.9047 的 AUC 值，证明了其在检测内部威胁方面的高效性和准确性，为组织提供了一种新的技术手段以应对内部安全威胁<sup>[6]</sup>。S.Delgado 采用自组织地图（SOM）人工神经网络模型，对拉里奥哈国际大学（UNIR）2015 至 2019 年间的 1,709,189 条在线学生记录进行了深入的无监督聚类分析，旨在适应新冠期间教育模式的转变，为线上线下结合的学习环境中的学生和教师提供更加个性化的支持，分析结果不仅揭示了特定用户群体与大学入学情况之间的关联，还发现了用户交互与学习绩效之间的显著正相关性，同时研究进一步表明，通过超越传统的技术导向方法，无论是在面对面还是在线学习环境中，都可以观察到相似的用户行为模式，这些发现为理解和优化在线学习体验提供了宝贵的洞见，并有助于教育工作者更好地支持学生的学习发展<sup>[7]</sup>。Mashhour 提出了一种针对管理信息系统中合法用户的非法操作和非法用户伪装进入系统问题的异常检测方案，通过从 Web 日志和点击流中获取用户行为数据，并将其分为角色属性和行为习惯两个维度进行建模，构建出用户的角色行为画像和习惯行为画像，进而融合这两种行为画像形成全面的用户行为模型，为模式匹配和异常行为实时检测的基础，从而完成对用户异常行为实时检测系统的开发<sup>[8]</sup>。

### 1.2.2 国内现状

虽然国内有关校园网用户行为分析的研究较少，但是经过多年的探索，逐渐形成了较为完善的分析体系。其中马仕玉舍弃传统 K-mediods 算法随机选取 k 个聚类中心的做法，通过选取 k 个彼此相对距离最远的初始聚类中心对 K-medios 算法提出改进，同时提出通过 PCA 分析降低多维数据维度以实现在输入层增加降维层和基于权值变化的训练停止条件的 SOM 改进算法，并将这两种算法应用于校园网用户行为分析中，获得了校园网用户上网行为规律。但是对于高维度数据其算法仍然具有较大的时间、空间复杂度<sup>[2]</sup>。因此宋坤引入基于图论的子空间聚类方法，结合线性惯性权重的粒子群

聚类算法，对于预处理后的数据进行了有效的降维，且一定程度上避免了局部最优解，得出有效的校园网用户网络行为模式<sup>[1]</sup>。马仕玉和宋坤主要集中在分析用户的网络使用情况，而朱峦通过联合 k 均值聚类和凝集聚类两种方法，通过凝集聚类来得到 k 均值聚类算法的初始聚类中心，提升了计算方法的剖析效力与聚类成果的平稳特性，通过对用户访问对象进行聚类，完成了以校园网为基础的客户拜访习惯方面的剖析，同时通过 Apriori 关联准则发掘出了用户的访问偏好<sup>[11]</sup>。贺雯静建立了完整的学生上网行为分析系统，从用户行为模式，学习成绩，学生用户群体，数据可视化与决策等方面建立了完善的分析体系，为校园网以及学生用户的管理提供了更便捷高效的解决方案<sup>[3]</sup>。祁家祯通过 Kafka 与 Flume 配合将传输到 Hive 数仓中，并使用 Spark 作为数据操作引擎将数据存储到 Mysql 中配合 web 页面的开发完成行为分析系统的开发，实现了大数据时代下海量用户数据的行为分析<sup>[11]</sup>。

### 1.3 研究内容

本文研究的主要内容如下：

①完成对校园网用户日志数据的预处理，将一天分为 8 个计数点，把一周的用户日志数据中的网络使用时间与流量使用情况转换到对应的计数点，通过将不同属性的数据标准化消除了尺度差异，最后通过加权归一化的方法实现不同属性融合分析。同时计算用户对应的网络协议使用次数以及流量使用情况，对这两种属性完成上述的预处理过程。

②对于预处理后的数据，比较 K-means++ 算法，DBSCAN 算法，以及自编码器的聚类效果，最终采用 K-Means++ 算法来完成对用户的聚类分析。初步实现对用户的上网时间喜好分析以及对于用户访问网站类型的偏好分析，帮助网络管理员更好的进行网络资源优化、安全管理、流量控制和用户支持，以提供更好的网络服务和用户体验。

③完成系统的需求分析和设计、用户访问协议可视化分析、各类用户的流量使用情况、上网时间的可视化分析以及用户偏爱的网站类型分析。

### 1.4 论文结构安排

本文共 6 章，每章的具体内容安排如下：

第 1 章是绪论。绪论部分介绍了本文的研究背景和意义，对校园网用户行为分析的国内外现状以时间发展为序进行了介绍，最后介绍了文本完成的工作。

第 2 章介绍相关理论与技术。本章详细阐述了数据预处理的各个环节，如数据清洗、转换、标准化和加权归一化。同时，深入介绍了本文所采用的三种聚类算法，并概述了系统开发所依赖的框架和技术。

第 3 章是预处理及聚类算法实现。依据第二章的预处理流程对本文数据集完成了全部的预处理，同时对三种聚类算法的效果进行了比较，使用聚类算法对用户的网络依赖程度和访问网站类型偏好进行了分析。

第 4 章是系统需求与设计。针对校园网用户行为分析提出了系统所需要完成的功能性需求，同时从实用性、可扩展性、离线处置的可连续特性和安全性对系统进行了非功能性分析。然后依据需求分析完成了对系统功能设计以及数据库相应表格的设计。

第 5 章是系统实现。根据对系统的需求分析和设计完成了对系统的开发，对系统要实现的各模块进行了展示。

第 6 章是系统测试。介绍了系统测试的目的，阐释了测试的相关方法，如黑盒测试和白盒测试，然后通过黑盒测试对系统的各模块的功能进行了测试。

最后通过总结与展望对本文的研究内容与工作进行了总结，指出了本文的不足的同时提出了未来对系统的改进方向。

## 1.5 本章小结

本章首先介绍了校园网用户行为分析的研究背景和意义，指出了互联网普及对人们生活方式和工作方式的深远影响，以及校园网络在教育领域的重要性。接着，概述了国内外在校园网用户行为分析方面的研究现状，包括国外在虚拟校园、在线学习平台、异常检测等方面的研究进展，以及国内在聚类算法、用户行为模式分析等方面的探索。明确了本研究的主要内容，包括校园网用户日志数据的预处理、聚类算法的比较与选择、系统需求分析与设计、系统实现与测试等。

最后，对论文的结构安排进行了说明，包括各章节的具体内容和研究工作的总结与展望。

## 第 2 章 相关理论与技术

### 2.1 预处理方法

数据预处理是为了让数据在经过清洗、变化、整合和优化的处理之后，使其转换为我们研究所需要的形式。通过进行数据预处理，可以改善我们数据的可信度、精确度和一致性，降低噪声和干扰因素的影响，并为后续的研究工作奠定良好的数据基础<sup>[13]</sup>。本文所采取的数据预处理流程如图 2.1 所示：

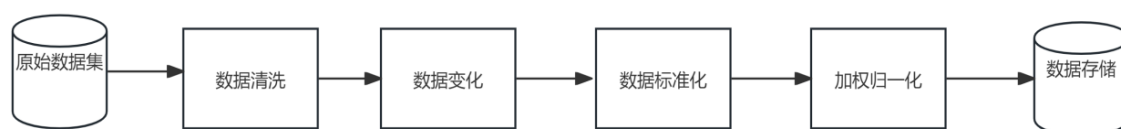


图 2.1 预处理流程图

#### 2.1.1 数据清洗

我们进行数据分析时，原始数据集难免会存在数据值缺失、错误、存在异常值和重复值等问题。数据清洗就是为了解决以上问题从而提高数据质量、确保分析的准确性和一致性，以及处理缺失值和确保数据的一体性，从而支持研究目的和论文的可信度。

#### 2.1.2 数据变换

经过数据清洗后的数据并不是所有的数据都属用户需要的形式，例如不同属性的数据值差距较大，又或者数据集并不适合我们后续所使用的分析方法，我们需要将数据集转换成适合的形式。

#### 2.1.3 数据标准化

由于原始数据不同属性的值具有不同的量纲和不同数据量级，所以为了使这些数据能够相互比较，需要进行数据标准化将数据按照比例进行缩放到相同的区间以达到去除量纲，便于不同量纲的数据进行比较和加权。

##### （1）Min-Max 标准化（Min-Max normalization）

Min-Max 标准化也叫做离差标准化，通过对原始数据进行线性变换将其映射到新的值域 $[\min, \max]$ 内，Min-Max 的标准化公式如下：

$$Y_i = \left( \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \right) \times (\max - \min) + \min, (1 \leq i \leq n) \quad (2.1)$$

公式 (2.1) 中对于序列  $X_1, X_2, X_3, \dots, X_n$  进行变换，得到新的序列  $Y_1, Y_2, Y_3, \dots, Y_n$  其中  $X_{\min}$  是原序列中的最小值， $X_{\max}$  是原序列中的最大值， $X_i$  是我们需要进行变化的数值，通过计算  $X_i$  在原始区间段所占的比率乘新区间的长度加上新区间的最小值得到变换后的值  $Y_i$ 。离差标准化的计算过程简单明了，但是对数据集变更敏感，每次变更后都需要重新计算最大最小值，同时离群点可能会扭曲整个数据集的分布。

### (2) Z-score 标准化 (zero-mean normalization)

Z-Score 标准化，也称为标准差标准化或零均值标准化。该方法是一种将数据转换为近似标准正态分布的办法，转化后的数据均值为 0，标准差为 1。这种转换首先会计算序列  $x_1, x_2, x_3, \dots, x_n$  的均值和标准差，然后通过减去数据集的均值并除以标准差来得到新的序列  $y_1, y_2, y_3, \dots, y_n$ 。Z-Score 标准化的公式如下：

$$y_i = \frac{x_i - \bar{x}}{s}, 1 \leq i \leq n \quad (2.2)$$

其中  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ， $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ ，分别代表原序列的均值和标准差。Z-score

适用于不知道数据系列中的最大值和最小值，或者存在噪声点的数据以及数据分布相对密集，且需要进行统计分析或建模的场景。

### (3) 例法 (归一化方法)

比例法 (归一化方法) 通常指的是将数据集中的每个特征值除以该特征的总和，使得变换后的特征值之和为 1。这种方法在处理概率分布、权重分配以及任何需要表示为比例或百分比的情况时非常有用。比例法归一化的一个常见形式是 L1 归一化，其公式如下：

$$y_i = \frac{x_i}{\sum_{i=1}^n x_i}, 1 \leq i \leq n \quad (2.3)$$

该方法具有简单、保持了数据的线性比例、正态化的特点，但是如果原始数据中包含零或接近零的值，比例法可能会导致归一化后的数据中出现零或非常小的值，这可能会影响模型的性能。同时该方法仅适合正值序列<sup>[13]</sup>。

#### 2.1.4 数据加权归一化

在对校园用户的日志分析时，往往多个属性之间具有一定的关联性，例如用户上网时间和用户上网使用流量等，通过加权归一化将多个属性综合考虑在一个属性中可以更好地反映多个属性的重要性和贡献，而不是仅仅关注其中某一个或多个属性。假设有  $n$  个属性  $X = \{x_1, x_2, x_3, \dots, x_n\}$ ，每个属性的权重  $W = \{w_1, w_2, w_3, \dots, w_n\}$ ，其中  $w_i$  是属性  $x_i$  的权重，且  $\sum_{i=1}^n w_i = 1$ 。每个属性的数据经过标准化处理后  $X' = \{x'_1, x'_2, x'_3, \dots, x'_n\}$ 。则加权归一化后的综合得分  $S$  如公式(2.4) 所示：

$$S = w_1 \cdot x'_1 + w_2 \cdot x'_2 + w_3 \cdot x'_3 + \dots + w_n \cdot x'_n \quad (2.4)$$

## 2.2 聚类算法介绍

聚类作为无监督学习领域的核心方法之一，其目标是根据数据点之间的相似性将它们划分为不同的组。这一过程确保了同一组内的数据点具有较高的相似性，而不同组之间的数据点相似性较低<sup>[14]</sup>。聚类分析的目的是发现数据内在的结构，它在许多领域都有应用，包括但不限于市场细分、社交网络分析、图像分割、基因表达分析等。常见的聚类算法如图 2.1：



图 2.1 常见的聚类算法图

本文采取基于划分的聚类算法中的 K-means 的改进算法 K-means++ 算法、基于密度算法中的 DBSCAN 算法以及基于模型聚类算法中的自编码器聚类算法来对经过预处理的校园网用户数据进行聚类分析。接下来，将对这三种算法进行详尽的阐述。

### 2.2.1 基于划分的聚类算法

基于划分的聚类算法通过迭代更新聚类中心，旨在实现数据集中的点在各自簇内高度相似，而簇间差异显著。这种方法的优势在于其较低的时间和空间复杂度，使其成为处理大规模数据集的有力工具。特别适合于在特征空间中呈现球状分布的簇的识别<sup>[13]</sup>。针对传统 K-means 算法对初始聚类中心敏感的特点，本文采取其改进算法 K-means++ 进行后续的分析<sup>[15]</sup>。K-means++ 算法的主要步骤如下：

- (1) 随机选择初始数据集中的点作为初始聚类中心。

- (2) 计算数据集中每一个点与最近的已选择的聚类中心的距离的平方  $D_{xi}$ ，然后使用轮盘赌的方法即  $P(x) = \frac{D_{xi}^2}{\sum_{x_j \in D} D_{xi}^2}$  来计算出下一个聚类中心。
- (3) 重复上述步骤直到得到我们所设定的  $k$  个聚类中心。
- (4) 将其余每个数据点划分到与它们距离最近的聚类中心。
- (5) 中心用质心的方法重新定位每类的中心。
- (6) 重复步骤④，⑤直到达到预设的迭代次数或聚类中心不再变化

### 2.2.2 基于密度的聚类算法

该类算法核心在于分析数据点的局部密度分布，并将密度较高的区域中的点归纳到同一簇中。基于密度的聚类算法的优点是其能够处理形态复杂的簇以及包含噪声的数据集，DBSCAN 算法作为基于密度聚类算法的代表，擅长辨识出不规则形状的簇。其基本流程如下：

- (1) 随机选择数据集中的—个数据点  $p$  作为起始点。
- (2) 识别出围绕数据点  $p$  的邻域内，满足特定密度条件的其他数据点集合。③重复上述步骤直到得到我们所设定的  $k$  个聚类中心。
- (3) 如果数据点  $p$  周围存在满足密度要求的点，则这些点与  $p$  一起构成一个密集区域，即形成一个簇。若未找到满足条件的点，则将  $p$  标记为边界点，并继续分析下一个数据点。
- (4) 持续执行密度探索和簇的形成步骤，直至遍历数据集中所有数据点，确保每个点都被归类为一个簇的成员或作为边界点。

### 2.2.3 基于自编码器的聚类算法

随着大数据时代的到临，数据量也迎来爆发式的增长，而传统聚类算法面临愈发复杂、非线性、高维度的数据表现不尽如人意，基于神经网络的聚类算法其具有更强的非线性建模能力和自适应性，可以处理复杂的数据分布和高维数据的特点，很好地弥补了传统聚类算法的缺陷。由于传统神经网络一般需要数据集具有显示的标签数据，但是无监督的数据并没有该类标签，因此自动编码器便被提出。假设数据集为  $x$ ，自编码器通过将数据集  $x$  作为监督信号来进行学习，最终实现映射  $f_\theta: x \rightarrow x'$ 。映射  $f_\theta$  可以拆分为两个过程  $g_{\theta_1}: x \rightarrow z$  和  $h_{\theta_2}: z \rightarrow x'$ ，分别对应编码器将高维数据编码为降维的压缩表示  $z$  以及解码器将压缩表示  $z$  解码为原始维度  $x'$ 。其结构图如图 2.2 所示：



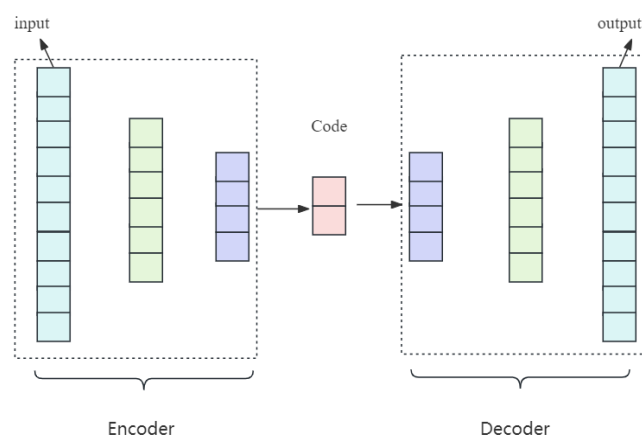


图 2.2 自编码器结构

自编码器的编码器有多个隐藏层组成，每个隐藏层包含多个神经元，隐藏层的神经元在接收到上层的输入数据后使用激活函数对其转换，最后一层隐藏层的输出即为经编码器降维后的特征向量 **Code**，其包含输入数据的关键信息。自编码器聚类便是将这降维后的特征数据再调用聚类算法进行聚类以实现对高维数据更好的聚类效果。其基本流程如图 2.3 所示：

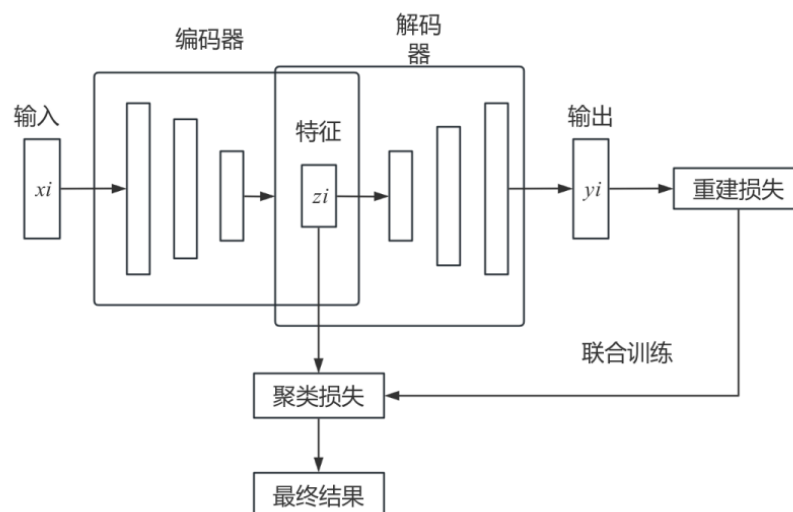


图 2.3 自编码器聚类流程

其中对自编码器的训练过程如下：

- （1）将输入数据传递给自编码器的编码器部分，获得编码后的特征表示。
- （2）将编码后的特征表示传递给自编码器的解码器部分，获得重构的输出。
- （3）计算重构输出与原始输入之间的损失，使用损失函数（如均方误差）衡量重构的准确性。
- （4）使用反向传播算法计算梯度，并使用优化器更新自编码器模型的参数。
- （5）重复以上步骤，迭代训练过程，直到达到预定的训练轮数（epochs）或损失收敛。

## 2.3 系统开发相应技术

### 2.3.1 前端展示层：Vue.js

在本研究中，前端展示层采用了 Vue.js 框架，这是一种构建用户界面的渐进式 JavaScript 框架。Vue 以其响应式数据绑定和组件化开发模式而闻名，它允许开发者以一种声明式的方式构建复杂的页面。本系统利用 Vue.js 的灵活性和易用性，实现了一个交互性强、用户体验良好的前端界面，从而确保了数据展示的直观性和操作的便捷性。

### 2.3.2 后端服务层：Java Spring Boot

系统后端服务层的开发选择了 Java Spring Boot 框架。Spring Boot 是一种基于 Spring 框架之上的微服务框架，它简化了基于 Spring 的应用开发流程，使得开发者能够快速构建独立、生产级别的 Spring 应用。通过 Spring Boot 的自动配置、嵌入式服务器和无代码生成特性，本研究中的系统能够高效地处理后端业务逻辑，同时保证了服务的稳定性和可维护性。

### 2.3.2 脚本自动化：Script

为了提高系统运行的自动化水平和效率，本研究在多个环节引入了脚本（Script）技术。用于自动化执行常规任务，如数据预处理、系统部署和测试等。通过编写和集成脚本，我们能够减少重复性工作，加快开发流程，并提高系统的可靠性。此外，脚本的使用还为系统提供了灵活性，使其能够快速适应不同的运行环境和需求变化。

## 2.4 本章小结

本章主要完成了对于校园用户数据预处理时所需要用到的相关方法进行介绍，通过清洗、集成、转换、标准化和加权归一化消除了不同属性的去量纲化，实现了对密切相关的数据的融合分析。介绍了常规的聚类算法的种类以及本文所使用的三种聚类算法的详细实现过程。最后本章介绍了本文系统所使用的研发框架以及技术。

## 第 3 章 预处理及聚类算法实现

### 3.1 校园网用户行为分析背景

#### 3.1.1 数据获取

文所使用的是重庆交通大学 Dr.com 计费系统所采集的学生网络日志。Dr.com 计费系统是一个多功能的网络管理平台，它通过集成用户管理、策略定制、设备监控和数据查询等多种功能，为学校网络运营提供了一套全面的解决方案，旨在优化网络资源的使用效率，提升用户体验，并确保网络安全和计费的准确性。Dr.com 计费系统的工作流程如图 3.1 所示：

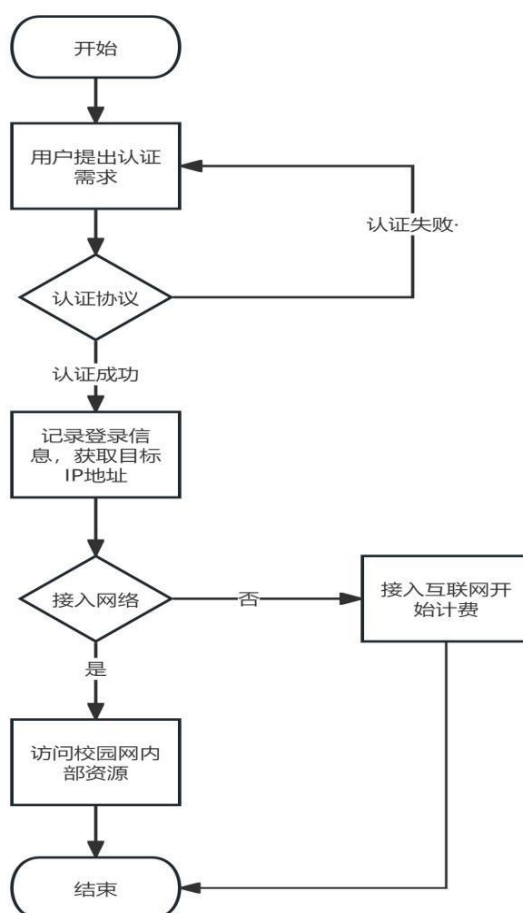


图 3.1 宽带认证流程图

### 3.1.2 数据特点

学校网络安全中心所提供的数据详细记录了“协议名称”、“起始时间”、“结束时间”、“上行流量”、“下行流量”、“用户账号”等 20 个属性。这些属性具有以下几个特点：

（1）数据内容详细。学生上网日志详细记录了用户的登录 ip、账号、访问网站的域名、ip、端口，单次上网所使用流量及时间等数据。确保了全方面、准确地对用户进行行为分析的可能性。

（2）数据真实准确。日志系统记准确录了用户的网络使用情况，反映了用户真实的网络活动。

（3）用户类型多样。日志系统所记录的用户包括教师、学校职工、学生等等，可以从各类用户入手分析其行为，本文仅分析学生用户行为。

（4）数据具有时序性。日志系统的每条数据都是严格遵循时间先后而记录的，所以具有时序性，本文数据变化便从这点入手。

（5）数据具有周期性。通常一周中，如果没有特别的情况，学生用户的行为节奏相对固定。

## 3.2 数据预处理

### 3.2.1 日志数据清洗

由于原始数据具有多个属性，并非每个属性都具有研究价值所以这里仅保留“协议”，“协议名称”，“起始时间”，“结束时间”，“上行流量”，“下行流量”，“用户账号”八个属性用作后续研究。同时对于缺失值、重复值、非学生用户、单次上网时间小于 1 分钟，单次上网流量使用小于 1Mb 的用户日志记录采取删除操作。本文分别对用户的上网规律以及用户对访问网站的偏好进行聚类分析，经过数据清洗后分别得到上网时间及流量和网络协议使用情况两个数据集。

### 3.2.2 日志数据变换

（1）上网时间及流量数据转换

已经清洗后的数据并不适合直接用来数据挖掘，结合用户上网的使用时间和使用流量两个特征具有的时序性，将数据转换为时间计数点所对应的值。这样做既保留了原始数据的时间特性同时还将数据转换成了适合我们所需要的形式，便于后续对学生上网规律的研究。具体的转换方法如下：

① 以 3 个小时为一个计数点将一天分为 8 个计数点，则一个周便有 68 个计数点。以这样的分割方式既能保证原始数据的时间特性不被破坏，同时快速有序地将原始时间格式的数据转换为计数点对应的值，便于更好地发掘用户上网的规律和对网络的依赖程度。

② 根据单次用户的日志记录计算对应时间点的值，记用户上线时间对应的计数点为  $\text{Login\_L}$ 、上线时间在该点未包括的值为  $\log in$ ；用户下线时间对应的计数点为  $\text{Logout\_R}$ 、下线时间在该计数点包括的时间为  $\log out$ ，该段时间包含的计数点为  $t_i$ ，具体的转化公式如下（时间以分钟为单位）：

$$value\_time = \begin{cases} 180 - \log in & (t_i = \log in\_L) \\ 180 & (\log in\_L \leq t_i \leq \log out\_R) \\ \log out & (t_i = \log out\_R) \end{cases} \quad (3.1)$$

该转换公式是时间转换公式，以其中一条用户数据进行举例说明，用户数据如下表所示：

表 3.1 用户上网时间及流量记录表

账号	上线时间	下线时间	使用流量/M
632*****17	2023/12/1 7:05:09	2023/12/1 10:53:15	229

根据时间转换公式该用户对应的时间计数点和值如表 3.2 所示：

表 3.2 用户单次上网时间转换表

	0	1	2	3	.....	248
632**** **17	0	0	115	113	0	0

对于流量的转换公式，首先需要计算每分钟用户所使用的平均流量，如公式（5-2）所示：

$$avg = \frac{flow}{logut\_time - logintime} \quad (3.2)$$

所以流量的转换公式如公式（5-3）所示：

$$value\_flow = value\_time \cdot avg \quad (3.3)$$

对应的根据网络转换公式该用户对应的时间计数点和流量的情况如表 3.3 所示：

表 3.2 用户单次上网流量转换表

	0	1	2	3	.....	68
632****	0	0	115.5	113.5	0	0
**17						

## （2）网络协议数据变换

由于原始日志数据对于用户主题类别进行了详细的划分，部分数据如图 3.2 所示：

协议	协议名称	起始时间	结束时间	i
udp	快手	2023/12/30 0:02	2023/12/30 0:02	1
udp	DNS	2023/12/30 0:05	2023/12/30 0:05	1
udp	Bilibili	2023/12/30 0:01	2023/12/30 0:02	1
udp	DNS	2023/12/30 0:05	2023/12/30 0:05	1
tcp	MSDS	2023/12/30 0:05	2023/12/30 0:05	1
udp	STUN	2023/12/30 0:02	2023/12/30 0:02	1
icmp	ICMP	2023/12/30 0:05	2023/12/30 0:05	1
udp	未知应用	2023/12/30 0:00	2023/12/30 0:02	1
udp	未知应用	2023/12/30 0:02	2023/12/30 0:02	1
tcp	其它下载	2023/12/30 0:04	2023/12/30 0:05	1
tcp	SYN_ACK	2023/12/30 0:05	2023/12/30 0:05	1

图 3.2 协议数据展示图

所以本文研究用户上网偏好从网络协议出发，统计了各协议的使用次数，以及流量使用情况，如图 3.3 所示：

name ▼	time	number
360更新	74	0
Android	83	10498789
Apex英雄	31	6626321
Apple/iCloud	440	192672314
BT扩展协议	183	44147
Bilibili	3703	296489521
Bittorrent	869	1272796
CCTV点播	45834	101887663
CIBN	92	39949574
DNS	21231	10114158
DOTA2/CSGO	250	360739976
EA游戏更新	38	449440
Epic游戏更新	335	659804351

图 3.3 协议数据变化后展示图

### 3.2.3 日志数据标准化

经过变换后的数据各属性间仍存在量纲不统一的问题，无法直接用于聚类算法研究。鉴于数据集稠密，且都为数值型数据的特点，为了消除不同属性间量纲的影响，分别对各属性使用公式（2.2）的 Z-score 方法来进行标准化。

### 3.2.4 日志数据归一化

对于学生用户的校园网使用时间及流量采取公式（2.4）来进行加权归一化，这里由于仅分析时间和流量两个属性所以公式（2.4）只需取  $S = w_1 \cdot x'_1 + w_2 \cdot x'_2$ ，其中  $w_1 = w_2 = 0.5$ 。

## 3.3 聚类效果对比及结果分析

这里采用预处理后的学生用户的网络使用时间以及流量使用情况作为数据集来进行三种聚类算法效果的比对，通过分别计算三种算法的聚类轮廓系数，Calinski-Harabasz 指数来实现对三种聚类算法聚类效果的评价。



### 3.3.1 自编码器训练

在对三种聚类算法的聚类效果进行比较前需要对自编码器进行训练，这里设置训练的轮数为 50 次，每次抽取 1000 个数据，通过记录每次迭代的损失函数来判断模型的训练情况，损失函数变化如图 3.4 所示：

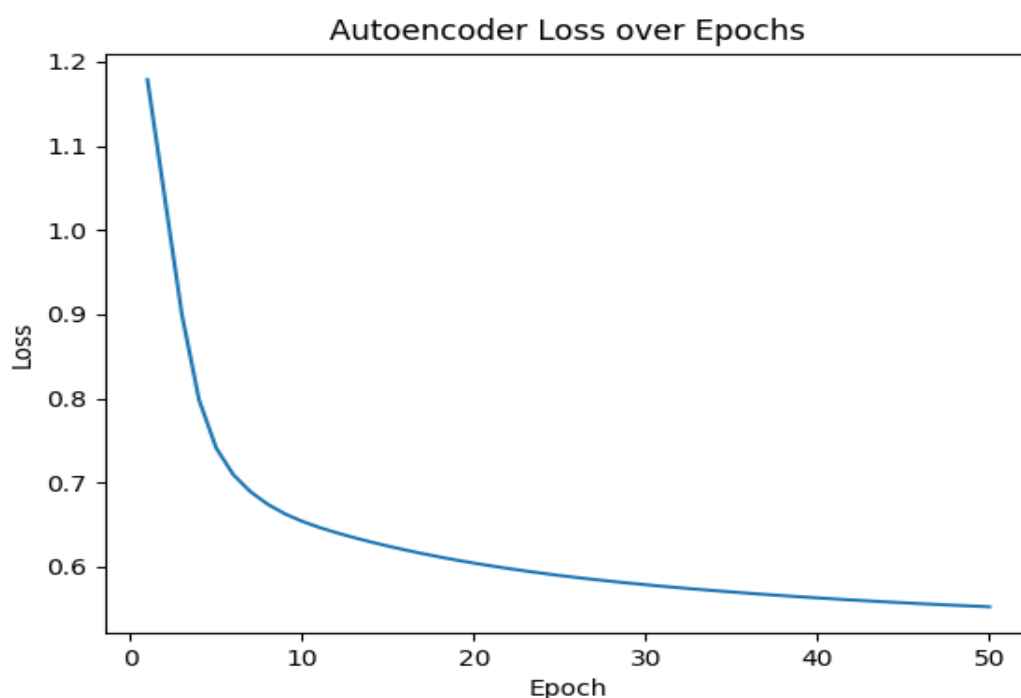


图 3.4 损失函数

通过图 3-3 可知损失函数逐渐收敛，模型训练完毕。

### 3.3.2 聚类效果对比

较大的轮廓系数通常被认为是更好的聚类结果，表示聚类效果较好，并且不同聚类之间的分离度较高。较小的轮廓系数可能意味着聚类结果不明显或不合理。Calinski-Harabasz 指数的取值范围是 $[0, +\infty]$ ，较大的指数表示较好的聚类结果，表示类间的方差相对较大，类内的方差相对较小。这意味着聚类之间的分离度高，聚类内部

的紧密度也相对较高。本次实验 K-means++ 聚类、DBSCAN 聚类、自编码器聚类的轮廓系数和 Calinski-Harabasz 指数分别如表 3.3 和表 3.4 所示：

表 3.3 聚类算法轮廓系数值表

聚类方法	轮廓系数
K-means++ 聚类	0.48871306185097696
DBSCAN 聚类	0.05457322790068288
自编码器聚类	0.45769022494017486

表 3.4 聚类算法 Calinski-Harabasz 指数数值表

聚类方法	指数数值
K-means++ 聚类	21820.0788008236
DBSCAN 聚类	275.858990894562
自编码器聚类	20491.01908357952

通过表 3.3 和表 3.4 可以知道对数据直接进行 K-means++ 聚类的效果差异不大，但两者效果都明显由于 DBSCAN 聚类方法，所以后续本文聚类分析采用 K-means++ 算法。

### 3.3.3 聚类结果分析

#### （1）用户网络依赖分析

采用 K-Means++ 算法对学生用户数据进行聚类分析，并通过肘部法则确定最优的簇数量。肘部法则是一种基于 WCSS（簇内平方和）与簇数量 K 之间关系的启发式策略，用以找到聚类分析中的最优簇数。WCSS 指的是每个簇的中心点与其成员点之间距离的平方和<sup>[16]</sup>。手肘法的步骤如下：

① 对不同的 K 值（聚类数），从 1 开始逐渐增加，每次使用相应的 K 值进行聚类分析。

② 对于每个 K 值，计算 WCSS，即将每个聚类中的所有点到该聚类中心的距离的平方和。

③ 绘制 WCSS 随 K 值增加的图表。

④ 在图表中寻找“手肘点”（Elbow Point），即 WCSS 随 K 值增加而减少速度变慢的点。这个点通常被认为是最佳的聚类数。

对归一化加权后的流量与上网时间数据使用手肘法，如图 3.5 所示，在  $k=5$  的时候出现拐点，所以这里取聚类数目为 5。

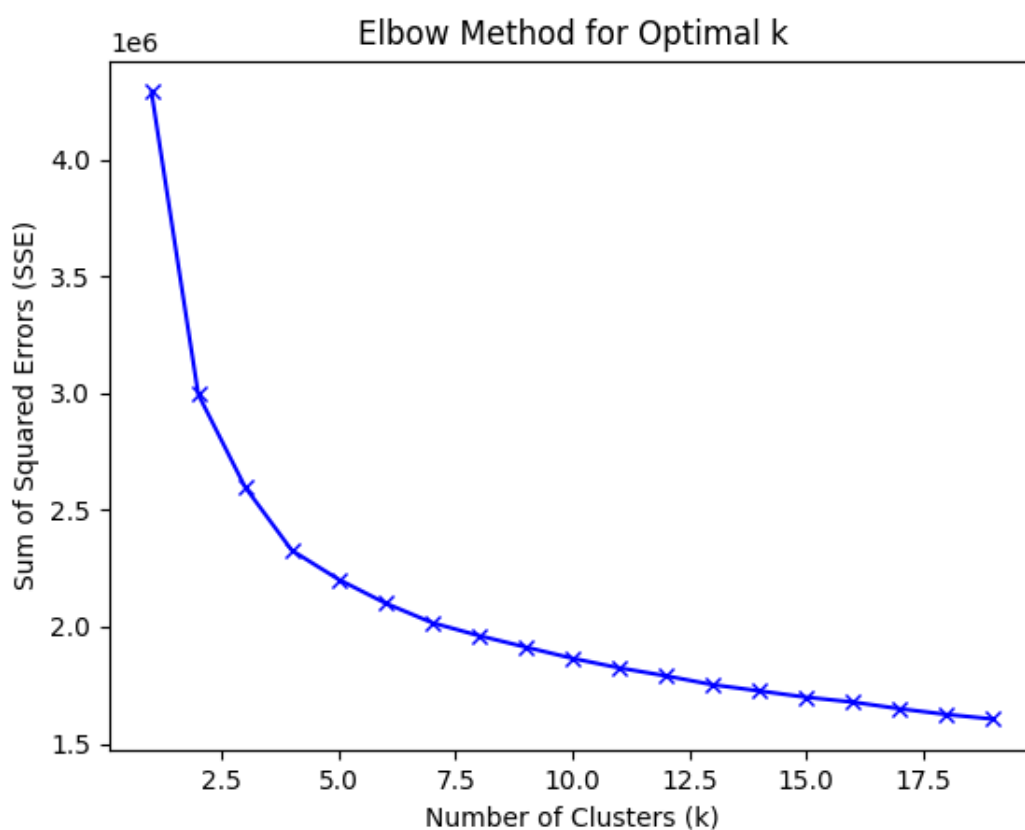


图 3.5 上网时间集流量聚类手肘图

根据聚类结果分别统计各类别在计数点上的平均使用网络时间以及平均使用流量情况，分别如图 3.6 和图 3.7 所示：

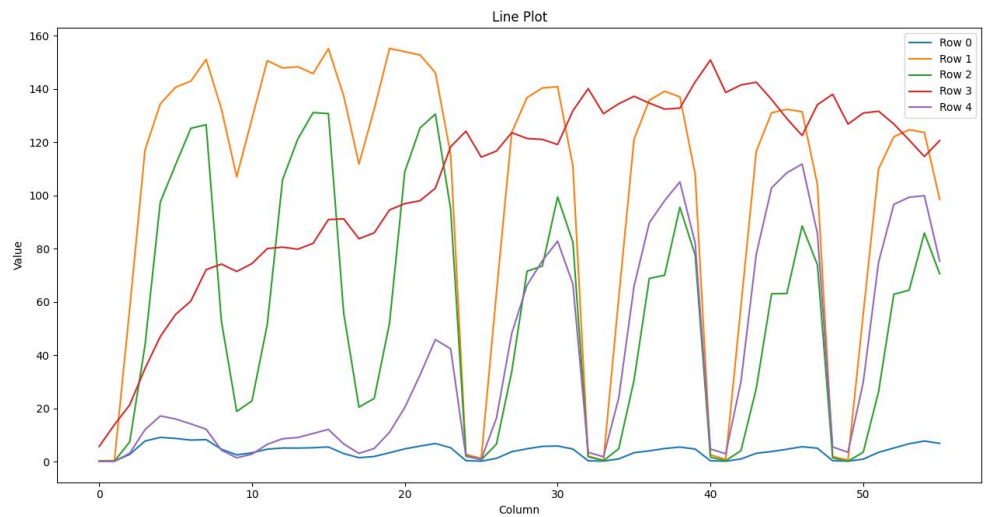


图 3.6 用户各节点网络使用时间情况

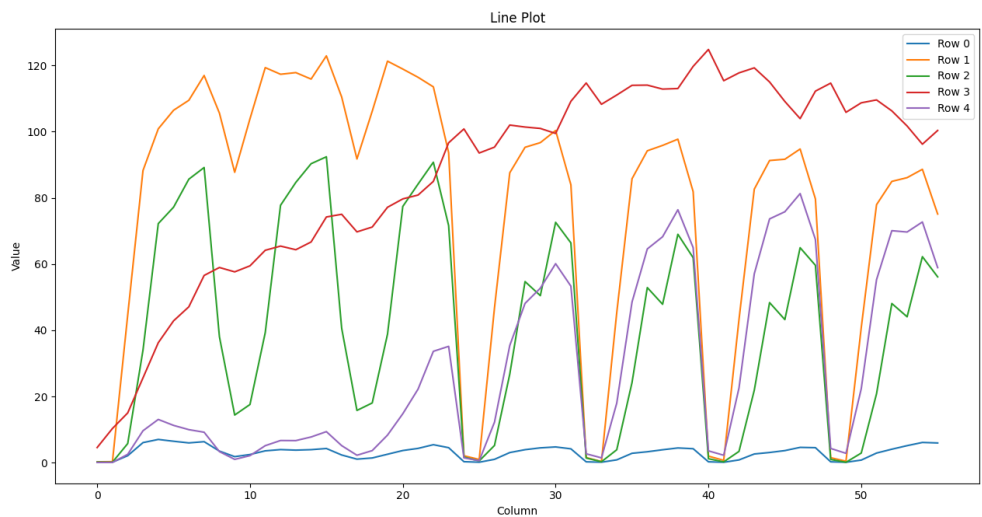


图 3.7 用户各节点网络使用流量情况

从聚类结果图可以得出如下规律：

首先从计数 0 到计数点 23 代表的时间段是周五、周六、周日，从计数点 24 到计数点 55 代表的时间是周一至周四。我们可以看到每个类别的用户明显存在以天为单位

存在周期性规律,大多数用户上网高峰期在每天的下午 3 点到次日的凌晨 1 点钟左右,这和学校课程安排相关,大多数人的课程大都集中在晚上 5 点 40 前,大多数用户的上网时间和流量的使用高峰期集中在上述时间段。针对各个类别的用户进行分析,可以看到用户类别 0 低网络依赖用户,该类用户一周的流量使用,与上网时间都处于较低的水平。用户类别 1 每天的流量使用与在线市场都处于较高的状态,尤其是周五周六以及周日,达到高峰,且这三天大多数用户一直处于在线状态并没有离线,可以推测这类用户沉迷于网络冲浪游戏无法自拔,辅导员可以对这类用户进行提示劝导。针对用户类别 2,这类用户对于网络也较为依赖,但是这类用户相比于类别 1,在周末不眠不休,该类用户在周末作息还算正常,每天凌晨后便下线休息了。针对用户类别 3,该类用户一周流量与使用时间数据呈逐渐上升趋势,猜测这类用户一周在进行类似于选修课刷课等类似的操作。针对聚类类别 4,该类用户整体上网依赖程度并不太高,并且,周末的网络使用情况明显少于上学时的使用情况,猜测这类用户大多周末都外出游玩了。

## （2）用户访问偏好分析

同样的首先使用手肘法对预处理后的网络协议数据进行分析来选取聚类的数目结果如图 3.8 所示:

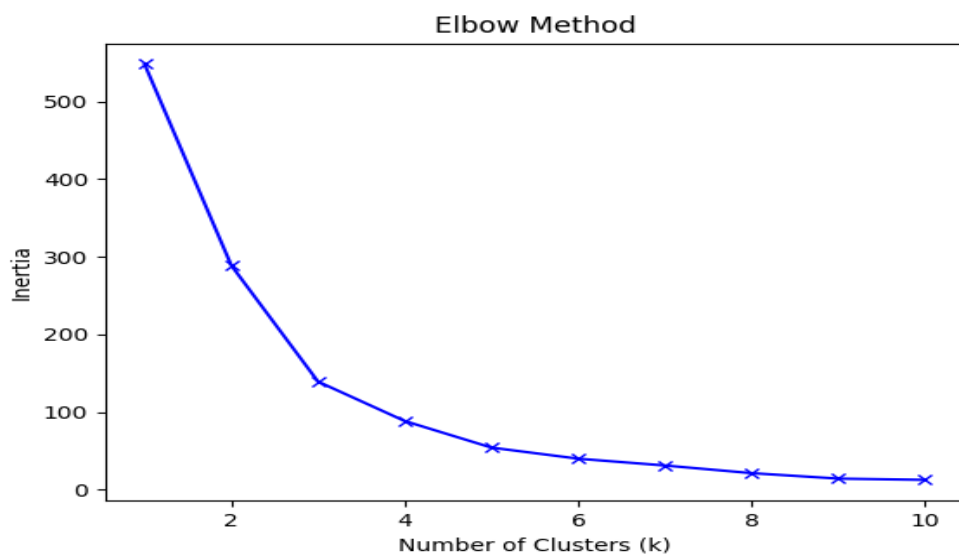


图 3.8 网络协议手肘图

根据手肘图我们可以看到在聚类数目达到 5 的时候出现拐点，所以这里设置 K-means++ 算法聚类数目为 5，对协议数据聚类后绘制其聚类结果的散点图，如图 3.9 所示：

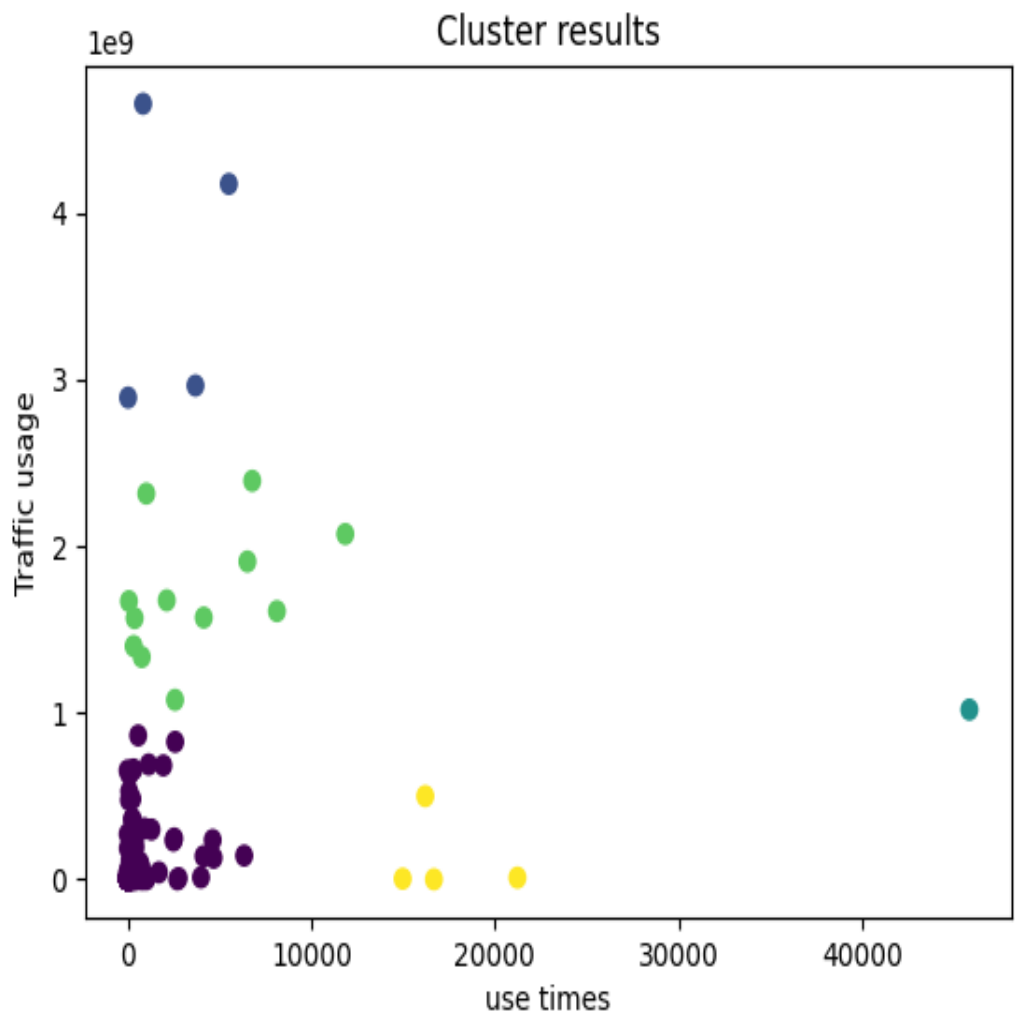


图 3.9 用户访问网站类型聚类散点图

根据聚类的散点图我们可以看到，依据用户访问网站所使用的网络协议的次数和使用流量我们大致可以将用户访问网站的类型分为五类，分别为低频次低流量使用类别，低频次中等流量使用类别，低频次高流量使用类别，中等频次低流量使用类别，高频次高流量使用类别，其中绝大多数的网络协议集中在左下角，而用户访问偏好更高的网站集中在较为稀疏的点中，我们将各簇结果展示出来，如表 3.5 所示：

表 3.5 聚类类别展示表

簇类别	簇大小	簇中主要成员
低频次低流量使用	253	WWW、微信聊天、快手、神之浩劫/无畏、淘宝天猫、其他游戏更新、腾讯、迅雷、百度
低频次中等流量使用	12	HTTP 分块传输、PPStream QQ 直播、QQ 聊天、优酷移动、华为应用商店、抖音歪歪语音/Bigolive 百度云盘、芒果 TV 等
低频次高流量使用	4	Bilibili、STEAM 游戏更新、微信语音视频、STUN
中等频次低流量使用	4	DNS、ICMP、SYN_ACK
高频次高流量	1	CCTV 播

由此总结出学生整体的访问偏好网站类型为：视频点播，音视频通信，网络游戏，网络购物，数据传输，音乐点播，搜索引擎，数据存储几类。

3.3 本章小结

本章介绍了校园网用户的日志数据来源于重庆交通大学的校园网日志系统，同时对获取的数据的特点进行了分析。通过第二章所介绍的数据预处理方法对我们的数据集完整了全部的预处理，比较了 K-means++聚类算法、DBSCAN 聚类算法、自编码器聚类算法的聚类效果同时选取了效果较好的 K-means++聚类算法，用手肘法确定不同数据集聚类数目来进行后续分析，完成了用户网络依赖分析以及用户访问偏好分析。



## 第 4 章 系统需求分析与设计

### 4.1 系统需求分析

#### 4.1.1 功能性需求分析

随着互联网的发展，对网络的使用已经成了我们生活密不可分的一部分。由于高校学生作为校园网用户的主要群体，本文拟通过对重庆交通大学的学生用户日志进行挖掘，得到校园网用户的行为特征以及总体的网络使用情况。辅助网络管理人员更好地监控校园网整体的运行使用情况，对网络的高效维护管理提供决策，同时更好地帮助辅导员完成对学生的在校管理，了解学生的网络依赖程度，以及学生对于不同类型网站的访问偏好情况。能够及时地对问题学生做出提醒。为了实现以上需求，校园网用户行为分析系统应该具有以下几方面的功能：

- （1） 用户登录管理界面；
- （2） 总体网络协议使用情况界面；
- （3） 用户网络依赖程度展示界面；
- （4） 用户访问偏好展示界面；
- （5） 日志基本信息管理界面。

#### 4.1.2 非功能性需求分析

该系统需要完成对校园网用户行为数据的挖掘分析，就得满足实际生活场景中的功能应用，保证系统良好的可扩展性，保证系统具有足够的经济价值以及安全性，因此系统应该满足以下的非功能性需求。

##### （1）实用性

校园网用户行为分析系统能够精确评估用户对网络的依赖程度，学生群体的访问偏好，优化网络资源分配，同时为校园网的安全管理、行为规范制定，以及提供定制化和决策支持提供数据基础，从而提升网络服务质量，增强网络安全防护，促进校园网络环境的健康发展。同时系统操作简单，数据展示明了，因此系统符合实用性的相关需求<sup>[17]</sup>。

##### （2）扩展性

本系统采用 Vue.js 和 Spring Boot 构建而成。这种架构允许前端界面和后端逻辑独立开发，实现前后端分离，使得在不干扰现有功能的情况下，可以轻松添加新特性或

进行维护升级。Vue.js 的组件化设计使得前端开发更加模块化，而 Spring Boot 的 RESTful 服务和微服务支持让后端能够灵活扩展。此外，Spring Boot 的依赖注入和插件机制进一步增强了系统的可维护性，而数据库层面的抽象则为未来可能的扩展或迁移提供了便利。

（3）离线处置的可连续特性

学生在校园网络中的活动频繁且更新速度快，随时间推移，累积的数据量日益庞大，这可能在系统处理历史数据时造成信息的冗余。本系统的目标是分析校园网用户群体的使用习惯和访问偏好，而非进行即时的在线分析。因此，系统需要能够支持我们处理逐渐收集到的信息，提炼出有价值的新内容，为决策调整提供支持。

（4）安全性

由于本系统涉及到学生的网络日志，属于高度隐私数据，所以本系统仅针对学校管理人员开放，其他无关人员无法获取传播学生相关的日志信息。

4.2 系统设计

4.2.1 系统开发环境

为了保障智慧社区信息服务平台的稳定开发，本信息服务平台采用相对稳定的开发环境进行开发，具体如表 4.1 所示：

表 4.1 系统开发环境表

开发工具	IntelliJ IDEA 2023.2.6、PyCharm Community Edition 2022.2.2
数据库	MySQL
浏览器	Microsoft Edge
操作系统	Windows 11

4.2.2 系统功能设计

系统的功能架构图如图 4.1 所示：

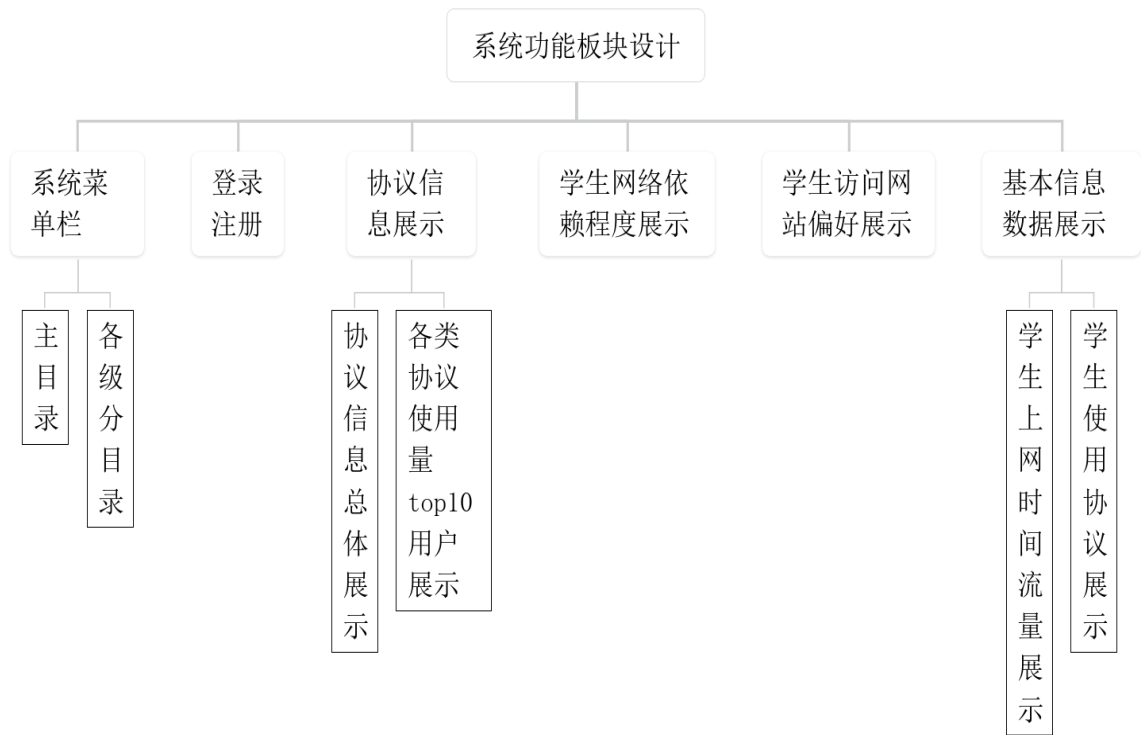


图 4.1 系统功能设计图

### 4.2.3 数据库设计

本论文所研究校园网用户行为分析系统的数据来自于重庆交通大学网络安全中心所提供的用户日志数据，由于原始数据集包含了用户的登录 ip，账号，访问网站的域名，ip，端口，单次上网所使用流量及时间等 20 个属性。且由于原始数据量庞大，我们需要删除对本文研究无关的数据字段，同时数据进行预处理。系统需要对处理完成后的数据进行存储并通过相应的前后端技术将数据在前端可视化出来，以方便系统用户进行相应的管理。所以我们需要设计数据库表格来存储我们相应的数据。相应的数据库表设计如下：

#### （1）网络协议类别表

网络协议表用于存储学生访问网站时所使用的相应的网络协议的基本信息，如网络协议名称、使用时间、使用流量、分类类别。具体表格设计如表 4.2 所示：

表 4.2 网络协议类别表

字段	含义	说明
name	网络协议名称	唯一主键，varchar(254)，用户每次上网所使用的网站名称的含义
time	使用时间	varchar(254)，用户每次上网所使用的时间（单位 min）
number	使用流量	varchar(254)，用户每次上网所使用的流量（单位 mb）
category	分类类别	varchar(254)，网站所属类别

（2）协议使用情况表

协议使用情况表用于存储学生每次使用网络所使用协议的基本信息，如协议名称、使用流量、用户 id，具体表的设计如表 4.3 所示：

表 4.3 协议使用情况表

字段	含义	说明
name	协议名称	varchar(254)，用户每次上网所使用名称的含义
number	使用流量	double，用户每次上网所使用的流量（单位 mb）
id	用户 id	唯一主键，varchar(254)，学生用户对应的账号

（3）各类用户平均上网时间情况表

各类用户平均上网时间情况表主要存储经过聚类算法处理后各类用户在相应平均上网时间的基本信息，如时间节点、分类类别、平均使用时间，具体表的设计如表 4.4 所示：

表 4.4 各类用户平均上网时间情况表

字段	含义	说明
time	时间节点	Int，以三小时作为一个时间节点，将一个月分为 248 个时间节点，从 0 开始计数到 247
category	分类类别	varchar(254)，用户每次上网所使用的流量（单位 mb）
data	平均使用时间	varchar(254)，各类别学生用户在对应时间节点上的平均上网时间（单位 min）

（4）各类用户平均上网流量情况表

各类用户平均上网时间情况表主要存储经过聚类算法处理后各类用户在相应平均上网流量的基本信息，如时间节点、分类类别、平均使用流量情况，具体表的设计如表 4.5 所示：

表 4.5 各类用户平均上网流量情况表

字段	含义	说明
time	时间节点	Int，以三小时作为一个时间节点，将一个月分为 248 个时间节点，从 0 开始计数到 247
category	分类类别	varchar(254)，用户每次上网所使用的流量（单位 mb）
data	平均使用流量	varchar(254)，各类别学生用户在对应时间节点上的平均上网流量（单位 mb）

（5）用户登录表

用户登录表主要存储登录用户的信息，如用户 id、用户姓名、用户账号、用户密码，具体的表的设计如表 4.6 所示：

表 4.6 用户登录表

字段	含义	说明
id	记录第 n 个用户的标签	自增，int，第 n 个注册的用户自动获得标签 n
name	用户名	varchar(254)，用户注册时输入的用户名
tel	用户账号	唯一主键，varchar(254)，用户注册时输入的电话号码，作为用户登录的账号
pass	用户密码	varchar(254)，用户注册后系统自动取电话号码后 6 位作为用户密码

（6）日志信息表

日志信息表主要存储用于用户网络依赖程度分析的日志数据基本信息，如用户账号、用户上线时间、用户下线时间、每次上网使用的流量、用户 ip 地址具体的表格设计如表 4.7 所示：

表 4.7 日志信息表

字段	含义	说明
account	用户账号	varchar(254)，第 n 个注册的用户自动获得标签 n
login	用户上线时间	varchar(254)，用户上线对应的时间
logout	用户下线时间	varchar(254)，用户下线时间对应的时间
flow	每次上网使用的流量	varchar(254)，用户注册后系统自动取电话号码后 6 位作为用户密码
ip	用户 ip 地址	varchar(254)，用户每次使用网络设备的 ip 地址

### 4.3 本章小结

本章结合实际情况提出了系统所需要完成的功能性需求，同时从实用性、可扩展性、离线处置的可连续特性和安全性对系统进行了非功能性分析。然后依据需求分析完成了对系统功能设计以及数据库相应表格的设计。

## 第 5 章 系统实现

### 5.1 登录模块

已注册用户在进入系统前需要先进行登录操作，用户需要输入用户名，密码，以及验证码，如图 5.1 所示：

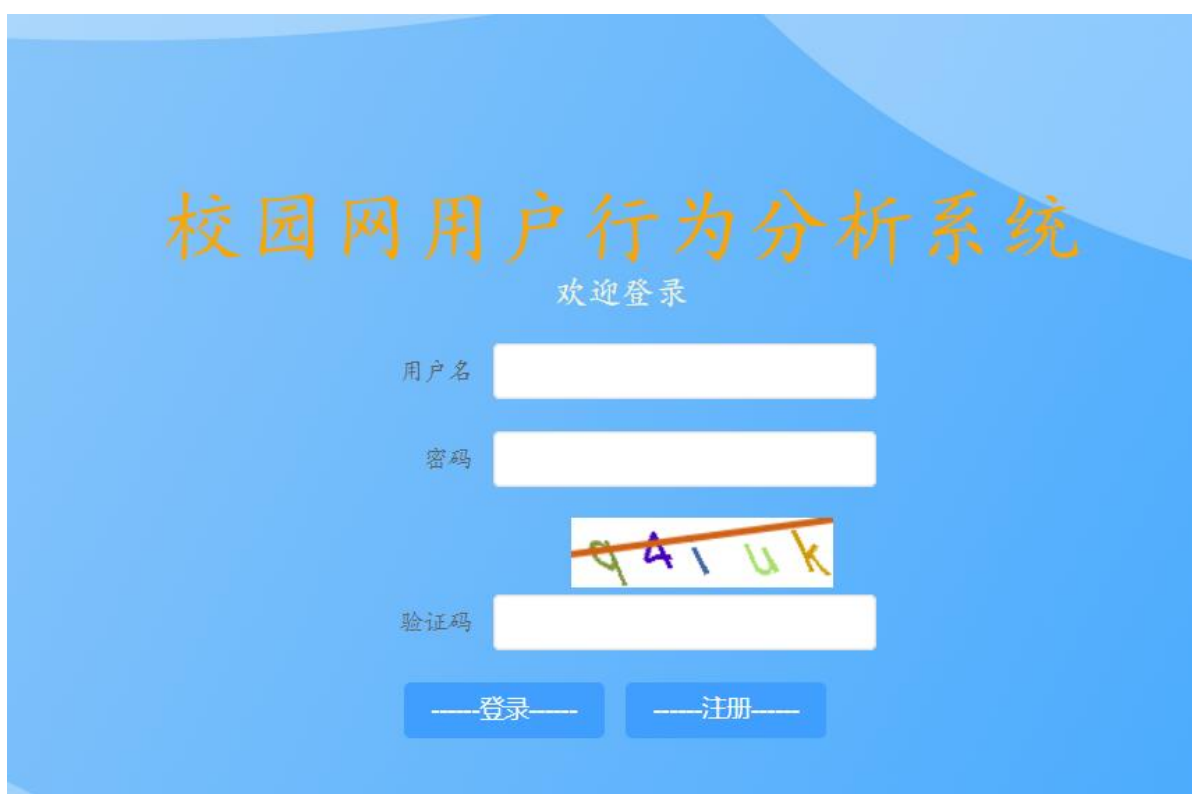


图 5.1 用户登录界面图

实现过程如图 5.2 所示：

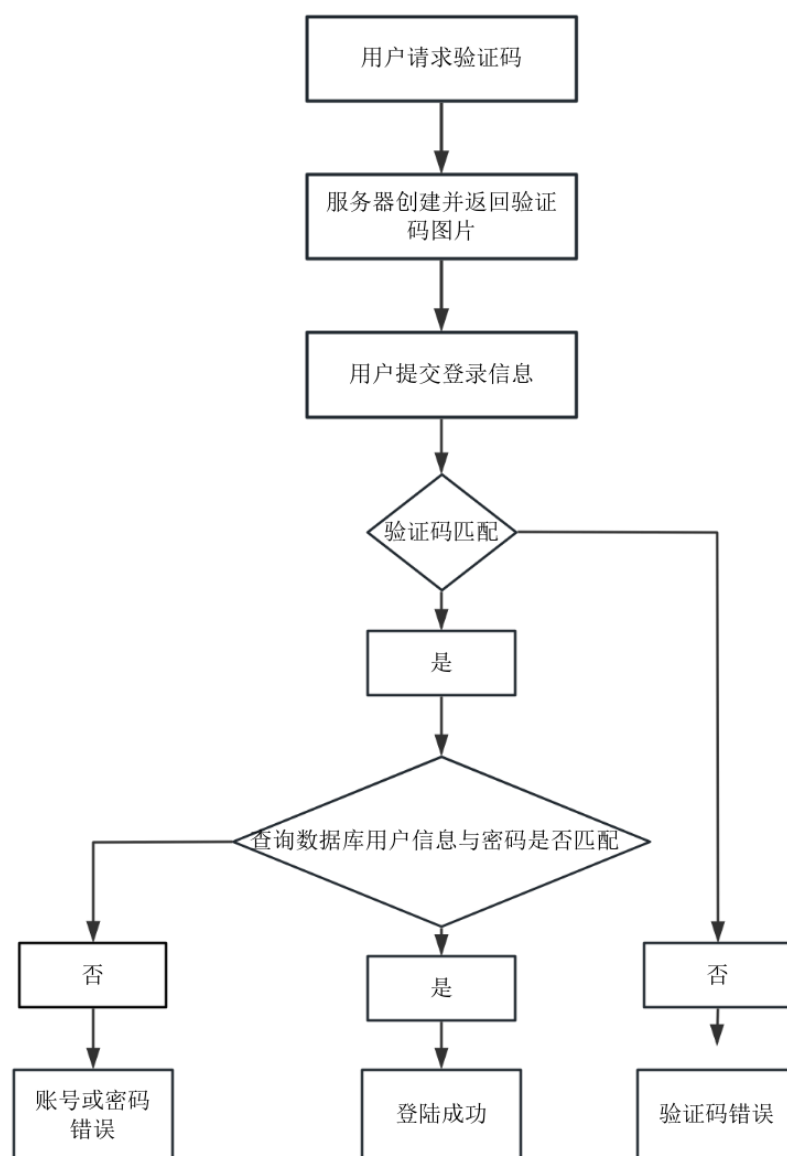


图 5.2 用户登录流程图

当我们想要获得二维码时，系统会通过/common/verify 生成并展示验证码图像，并在 session 中记录。提交登录时，用户需输入账号、密码及验证码，系统在/loginc 路径下先校验所输入的验证码，若验证码正确，则进一步验证账号密码。验证成功后，返回包含用户 ID 的 JSON 成功响应；验证码和账号密码任意一个验证错误会抛出对应的提示。



## 5.2 注册模块

未注册过的用户需要进行注册操作，新用户需要输入姓名和手机号来进行注册，系统会自动将手机号作为新用户的账号，手机号后 6 位作为新用户的密码，注册页面如图 5.3 所示：



The image shows a user registration page for a system titled '校园网用户行为分析系统' (Campus Network User Behavior Analysis System). The page has a blue background with a white title and subtitle '欢迎注册' (Welcome to Register). Below the title, there are two input fields: one for '姓名' (Name) and one for '手机号码' (Mobile Number). Below these fields is a blue button with the text '——注册——' (Register).

图 5.3 用户注册页面图

实现流程如图 5.4 所示：

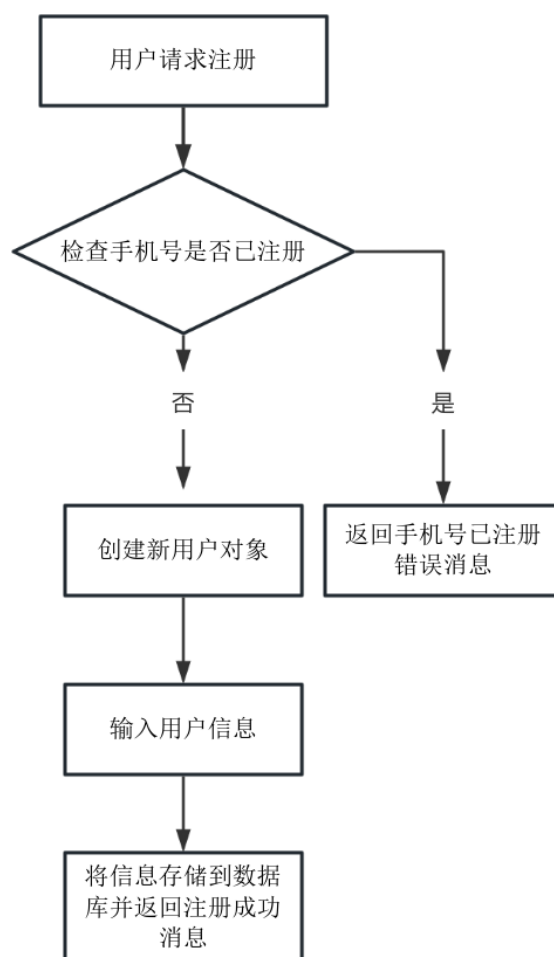


图 5.4 用户注册流程图

它通过 POST 请求接收用户的注册信息。系统首先验证用户输入的手机号是否唯一，若该手机号未被注册，服务端将使用手机号的后六位作为密码创建一个新用户账户，并将用户信息存储到数据库中，注册成功后返回成功消息；若手机号已存在，则返回错误信息提示用户手机号已被注册。

### 5.3 协议信息展示模块

系统会自动统计各类协议的使用次数以柱状图的形式统计，同时计算每个包所占比例，以饼图形式展示。对于每次网络使用的数据包大小进行统计将其归为五类，分别为小于 1mb，大于 1mb 小于 3mb，大于 3mb 小于 6mb，大于 6mb 小于 9mb 以及大

于 9mb，将数据以饼图形式显示。最后系统会统计每个协议使用量前 10 的用户，将结果以条形图显示出来。如图 5.5 至图 5.9 所示：

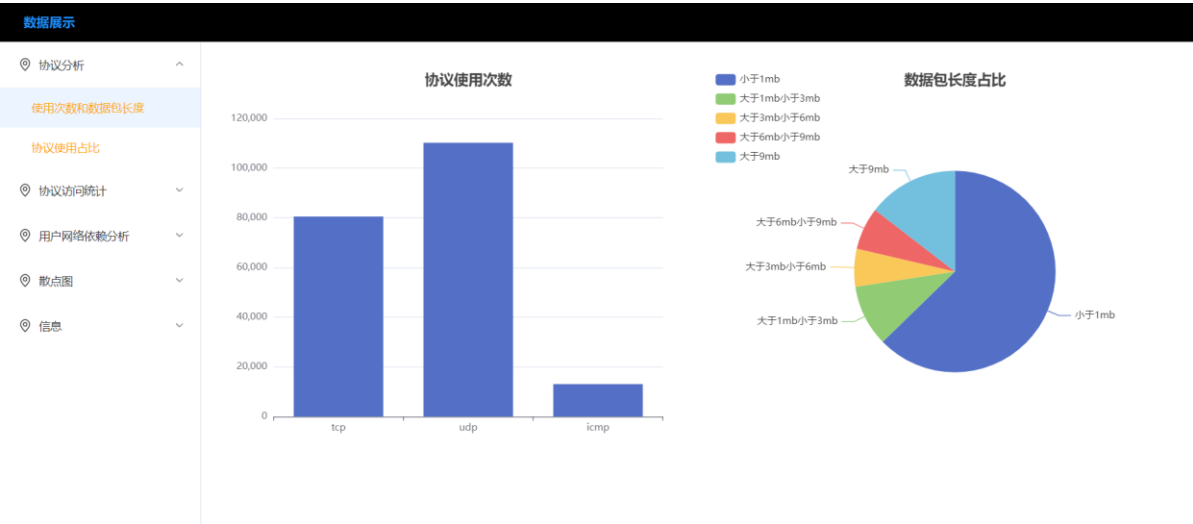


图 5.5 协议使用次数与总体数据包长度图

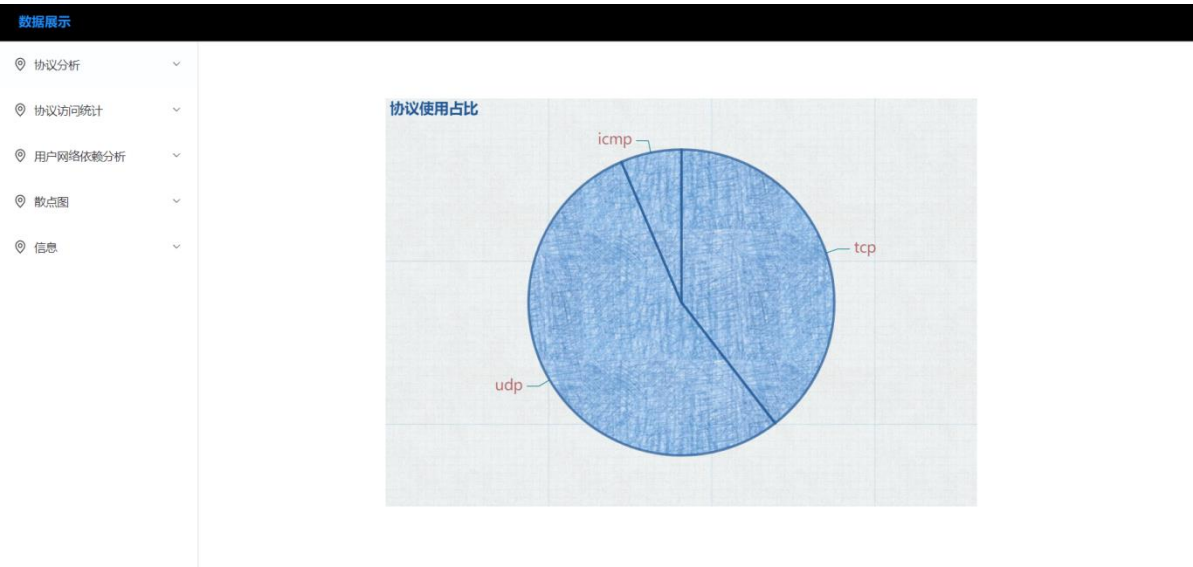


图 5.6 协议使用占比图

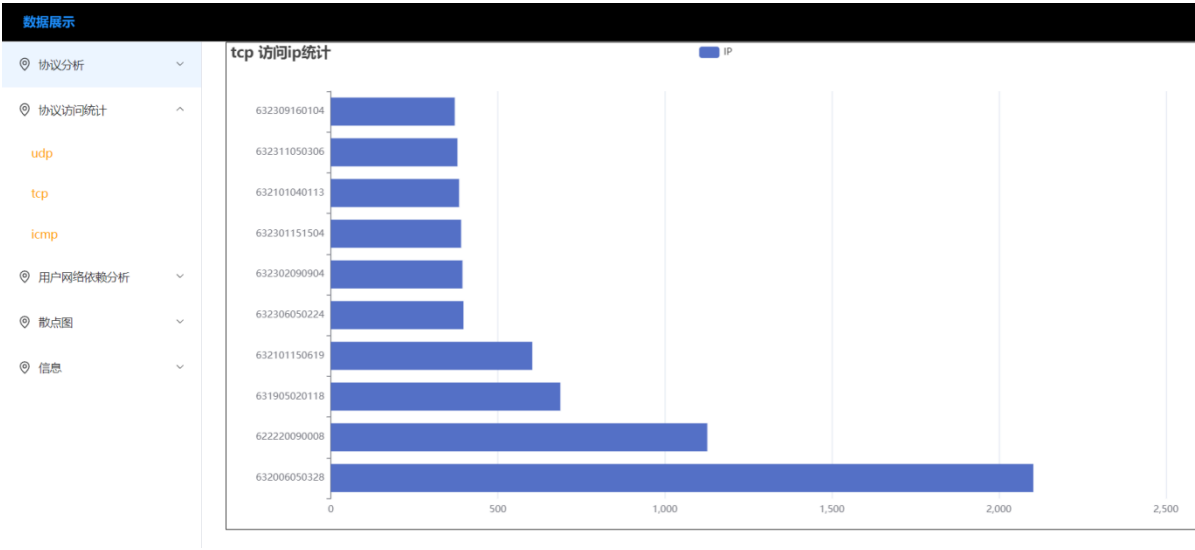


图 5.7 tcp 协议信息用户访问量 top10 图

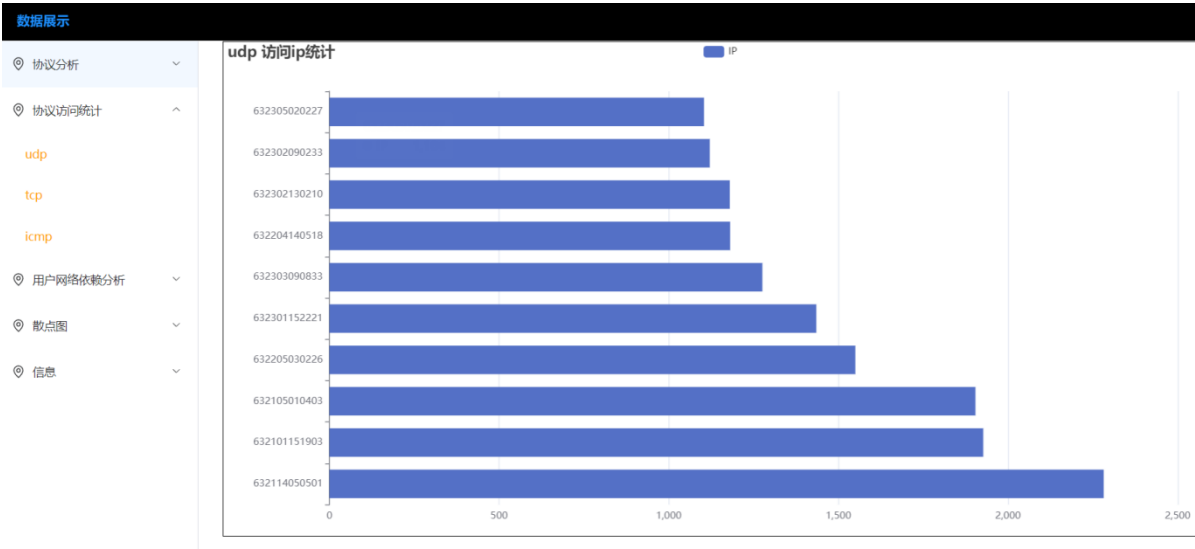


图 5.7 udp 协议信息用户访问量 top10 图

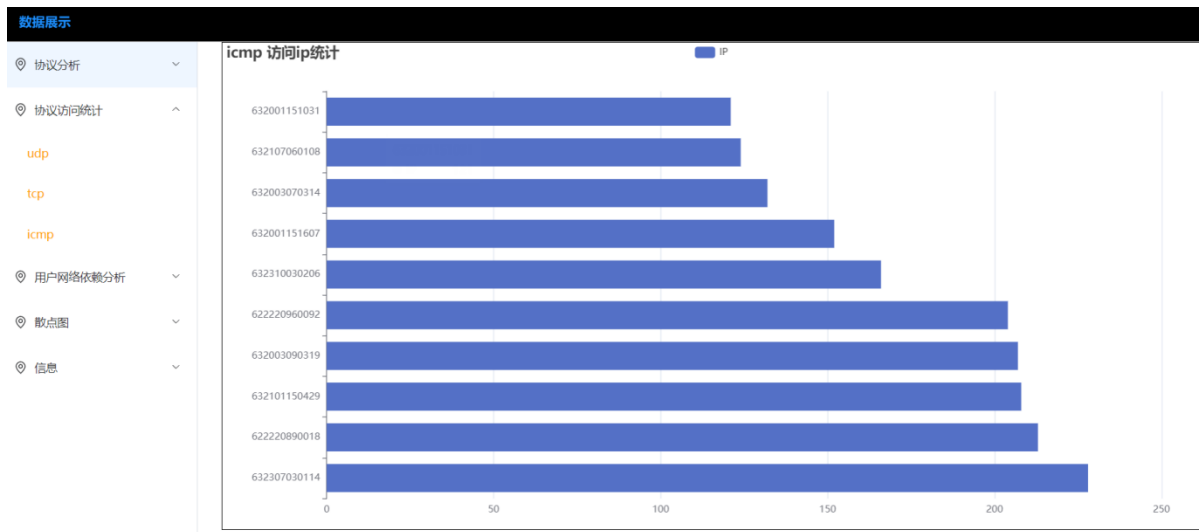


图 5.9 icmp 协议信息用户访问量 top10 图

## 5.4 用户网络依赖程度展示模块

用户可以清晰地看见经过聚类算法聚类后各类型的用户在各个时间节点对应的平均使用时间以及平均流量使用情况，便于辅导员对于成谜网络的学生做出提醒。用户也可以点击图片右上角的类别按钮选择需要展示的类别进行展示。如图 5.10 和图 5.11 所示：

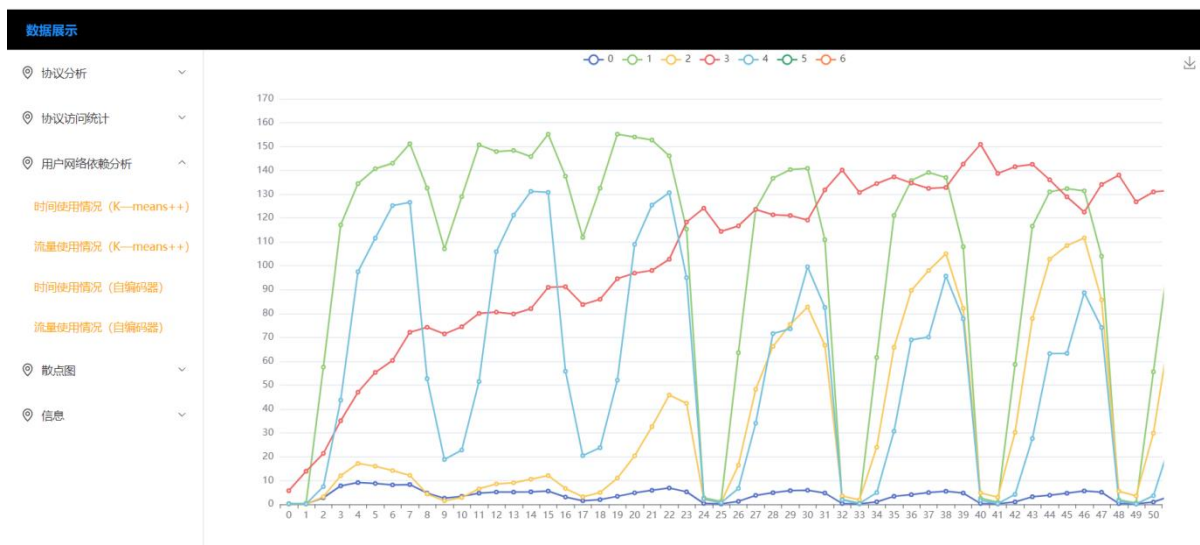


图 5.10 基于 K-means++ 的各类用户网络依赖程度展示图

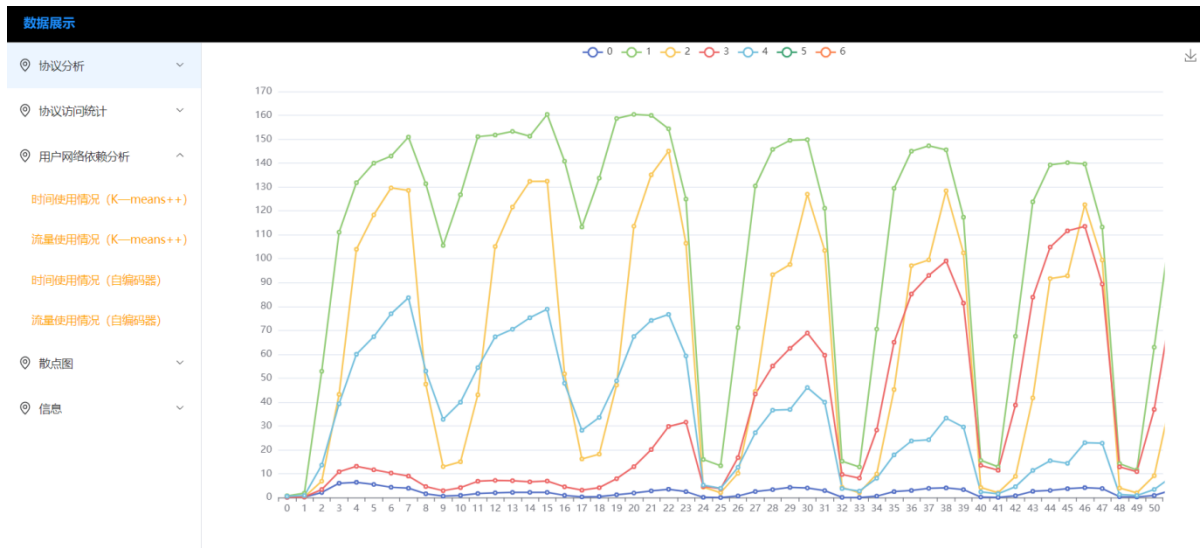


图 5.11 基于自编码器的各类用户网络依赖程度展示图

## 5.5 用户访问偏好展示模块

用户可以看到根据网络协议的访问次数和使用流量聚类后的图片结果，聚类结果共分为 5 类，用户可以选择右上角的按钮选择需要的类别进行展示。如图 5.12 所示：

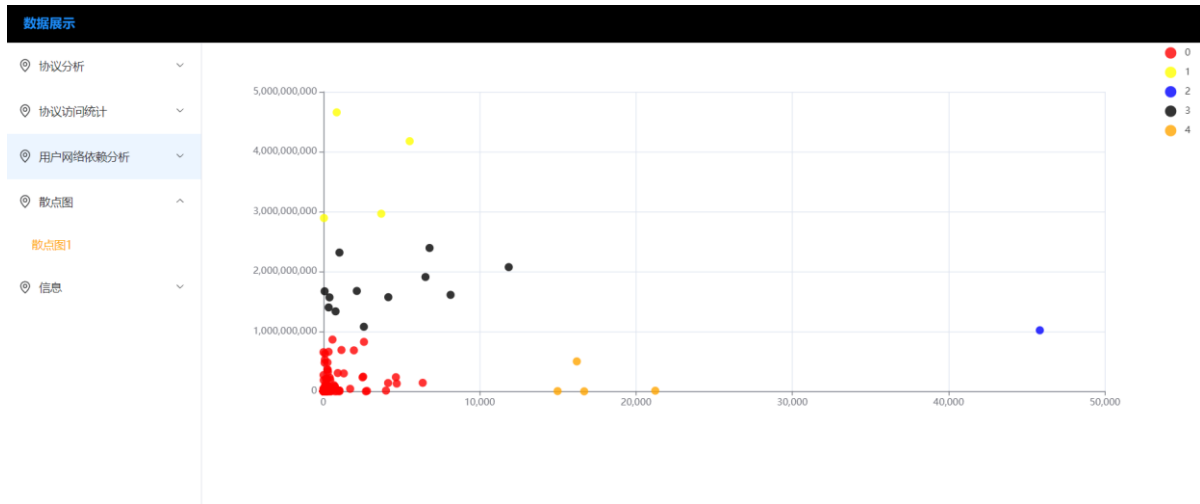


图 5.12 用户访问网站偏好展示图

### 5.6 基本信息展示模块

如图所示，在该模块中可以看到基本的日志信息，包括用户账号、每次访问所使用的协议类型以及数据包长度。用户可以通过完整学号来对学生进行精确的搜索，或者输入部分数字来进行模糊搜索。同时可以选择每页显示的数据条数，有每页 5 条，每页 10 条，以及每页 20 条的选项。同时也可以进行页数的跳转，浏览相应的数据，针对每条数据，我们的用户还可以进行相应的修改和删除操作。如图 5.13 所示：

数据展示

📍 协议分析

📍 协议访问统计

📍 用户网络依赖分析

📍 散点图

📍 信息

输入学号关键字

输入请求类型

查询

学号	类型	数据包长度	操作
632306050224	tcp	2.794921875	<a href="#">删除</a> <a href="#">修改</a>
632005070417	udp	1.4140625	<a href="#">删除</a> <a href="#">修改</a>
632262020302	udp	3.130859375	<a href="#">删除</a> <a href="#">修改</a>
632005070417	udp	1.4140625	<a href="#">删除</a> <a href="#">修改</a>
632001150626	udp	1.015625	<a href="#">删除</a> <a href="#">修改</a>
622220960092	icmp	0.15625	<a href="#">删除</a> <a href="#">修改</a>
632203070214	udp	380.1650390625	<a href="#">删除</a> <a href="#">修改</a>
632204140315	udp	0.3515625	<a href="#">删除</a> <a href="#">修改</a>
632006050330	tcp	0.7421875	<a href="#">删除</a> <a href="#">修改</a>
632362010211	udp	0.16015625	<a href="#">删除</a> <a href="#">修改</a>

共 203809 条

10条/页

< 1 2 3 4 5 6 ... 20381 >

前往

1

页

图 5.8 基本信息展示图

### 5.7 本章小结

本章主要完整了对校园网用户行为分析系统的开发，对其各功能模块进行了展示，同时介绍了较为复杂的功能模块的实现流程。

## 第 6 章 系统测试

### 6.1 系统测试的目的

系统测试是我们软件开发流程中必不可少的一个环节，通过对软件的全面测试，来检测是否每个功能都按照需求说明所执行，有效地帮助开发人员找出系统中存在的错误和缺陷。同时评估系统在各种条件下的性能，如他的响应时间，消耗的资源等等，确保我们的系统能以高性能稳定地运行下去。

### 6.2 系统测试的方法

目前，系统测试主要采用两种方法：黑盒测试和白盒测试。黑盒测试，也被称作功能测试或规格测试，它不依赖于软件的内部结构和实现细节<sup>[18]</sup>。在进行黑盒测试时，我们关注的是软件的功能是否符合需求规范，即软件是否按预期执行其功能。测试过程中，我们模拟用户操作，验证软件的输入、输出和处理流程是否正确。白盒测试，又称为结构测试或代码测试，与黑盒测试相对，它侧重于发现代码层面的问题，如逻辑错误、路径问题、资源泄露和性能瓶颈等。在白盒测试中，测试者需要对软件的内部结构有深入地了解<sup>[19]</sup>。对于校园网用户行为分析系统，我们采用的是黑盒测试方法，这意味着在测试过程中，我们不需要了解程序的具体实现代码。

### 6.3 系统测试实现

该功能模块的测试用例表如表 6.1 所示：

表 6.1 用户注册测试表

测试编号	1 号
测试类型	用户注册
前置条件	新用户填入相关信息进行注册
预期结果	注册后系统自动跳转到登录页面并显示注册成功，数据库成功写入新用户数据
实际结果	与预测结果一致

用户注册功能的测试结果展示如图 6.1 至图 6.4 所示：





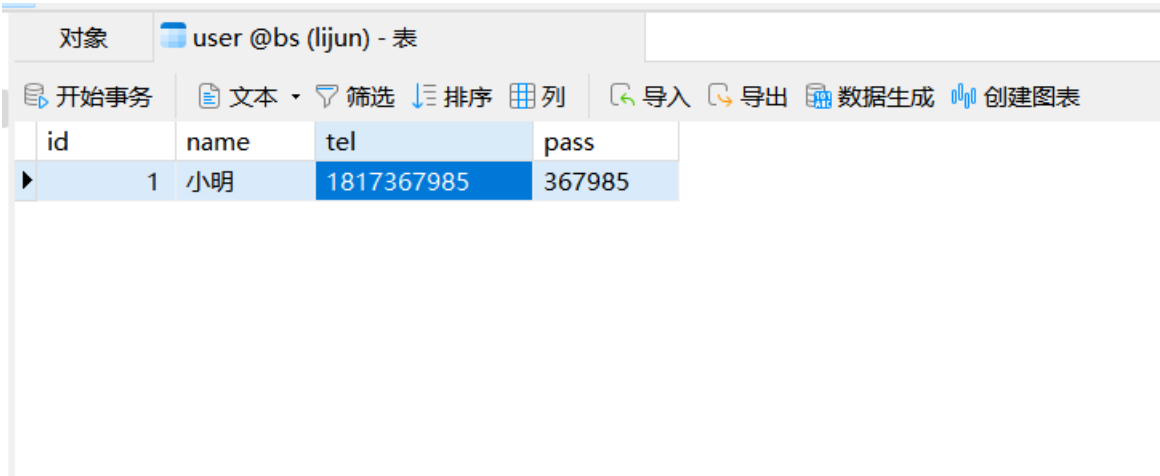
The image shows a registration form for a system titled '校园网用户行为分析系统' (Campus Network User Behavior Analysis System). The subtitle is '欢迎注册' (Welcome to Register). The form has a blue background. It contains two input fields: '姓名' (Name) with the value 'xiaoming' and '手机号码' (Mobile Number) with the value '130363829978'. Below these fields is a blue button with the text '——注册——' (Register).

图 6.1 用户注册信息填入图



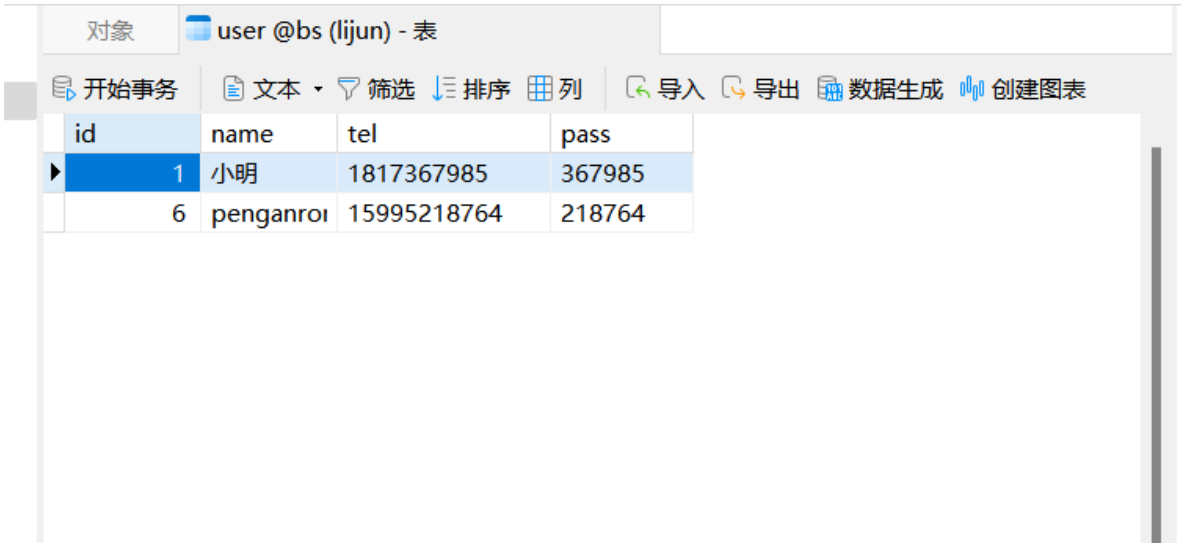
The image shows the login page of the '校园网用户行为分析系统' (Campus Network User Behavior Analysis System). The subtitle is '欢迎登录' (Welcome to Login). At the top, there is a green notification box with a checkmark and the text '注册成功默认密码为手机号后六位' (Registration successful, default password is the last six digits of the mobile number). The login form has three input fields: '用户名' (Username), '密码' (Password), and '验证码' (Verification Code). The verification code field shows a CAPTCHA image with the characters '941uk'. Below the input fields are two blue buttons: '——登录——' (Login) and '——注册——' (Register).

图 6.2 用户注册成功跳转登录页面图



对象	user @bs (lijun) - 表		
开始事务	文本	筛选	排序
列	导入	导出	数据生成
创建图表			
id	name	tel	pass
1	小明	1817367985	367985

图 6.3 新用户注册前数据库图



对象	user @bs (lijun) - 表		
开始事务	文本	筛选	排序
列	导入	导出	数据生成
创建图表			
id	name	tel	pass
1	小明	1817367985	367985
6	penganroi	15995218764	218764

图 6.4 新用户注册后数据库图

表 6.2 登录验证码错误测试表

测试编号	2 号
测试类型	用户登录
前置条件	用户登录时验证码输入错误
预期结果	系统自动提示验证码错误
实际结果	与预测结果一致

用户登录时输入的登录验证码错误的测试结果如图 6.5 所示：



图 6.5 验证码错误图

表 6.3 登录账号密码错误测试表

测试编号	3 号
测试类型	用户登录
前置条件	用户登录时验证码输入正确，但账号密码错误
预期结果	系统自动提示验证码或账号错误
实际结果	与预测结果一致

用户登录时验证码输入正确，但账号密码错误的测试结果展示如图 6.6 所示：



图 6.6 账号或密码错误图

表 6.4 用户登录输入正确测试表

测试编号	4 号
测试类型	用户登录
前置条件	用户登录时验证码输入正确，账号密码输入正确
预期结果	成功进入系统
实际结果	与预测结果一致

用户登录时验证码输入正确，账号密码输入正确的测试结果展示如图 6.7 所示：

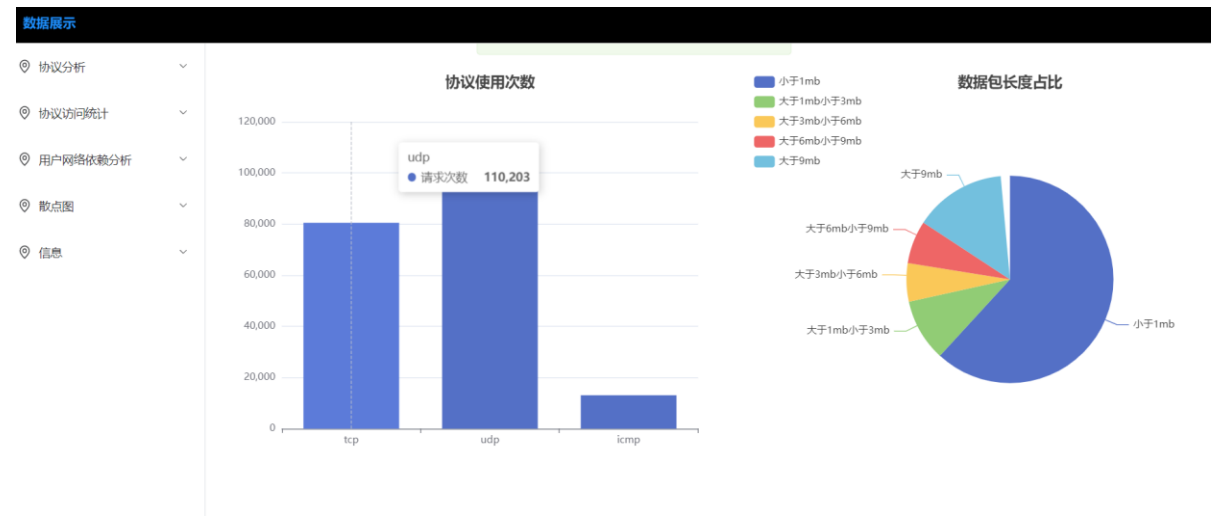


图 6.7 用户成功登录图

表 6.5 用户信息搜索测试表

测试编号	5 号
测试类型	信息搜索
前置条件	用户根据学号或者网络协议进行模糊搜索， 用户选择用户修改每页展示数据条数进行展示
预期结果	成功显示相应的用户信息，进行页面跳转， 页面显示的信息条数成功改变
实际结果	与预测结果一致

用户根据学号或者网络协议进行模糊搜索、用户选择修改每页展示数据条数进行展示的测试结果展示如图 6.8 至图 6.10 所示：

数据展示

📍 协议分析

⌵

📍 协议访问统计

⌵

📍 用户网络依赖分析

⌵

📍 散点图

⌵

📍 信息

⌵

632

t

查询

学号	类型	数据包长度	操作
632306050224	tcp	2.794921875	<a href="#">删除</a> <a href="#">修改</a>
632006050330	tcp	0.7421875	<a href="#">删除</a> <a href="#">修改</a>
632014050201	tcp	2.3251953125	<a href="#">删除</a> <a href="#">修改</a>
632101150606	tcp	2.537109375	<a href="#">删除</a> <a href="#">修改</a>
632306050224	tcp	2.794921875	<a href="#">删除</a> <a href="#">修改</a>
632004140303	tcp	0	<a href="#">删除</a> <a href="#">修改</a>
632302090305	tcp	5.826171875	<a href="#">删除</a> <a href="#">修改</a>
632009120420	tcp	0	<a href="#">删除</a> <a href="#">修改</a>
632208040615	tcp	4.85546875	<a href="#">删除</a> <a href="#">修改</a>
632004140303	tcp	0	<a href="#">删除</a> <a href="#">修改</a>

共 70049 条

10条/页

< 1 2 3 4 5 6 ... 7005 >

前往

1

页

图 6.8 用户信息模糊查询图

数据展示

📍 协议分析

⌵

📍 协议访问统计

⌵

📍 用户网络依赖分析

⌵

📍 散点图

⌵

📍 信息

⌵

632

t

查询

学号	类型	数据包长度	操作
632204141202	tcp	0.4296875	<a href="#">删除</a> <a href="#">修改</a>
632310040309	tcp	0	<a href="#">删除</a> <a href="#">修改</a>
632107110220	tcp	313.9345703125	<a href="#">删除</a> <a href="#">修改</a>
632208040115	tcp	0	<a href="#">删除</a> <a href="#">修改</a>
632005010320	tcp	0.650390625	<a href="#">删除</a> <a href="#">修改</a>

共 70049 条

5条/页

< 1 ... 998 999 1000 1001 1002 ... 14010 >

前往

1000

页

图 6.9 用户信息跳转查询图

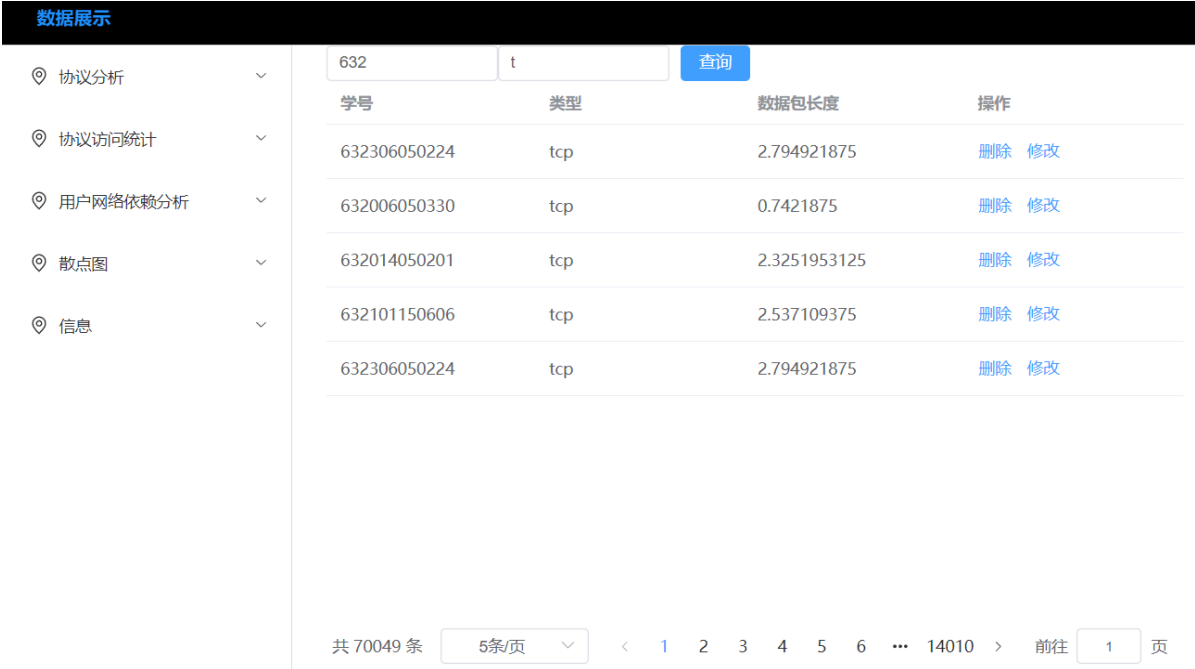


图 6.10 转变页面用户信息展示条数图

表 6.6 用户信息修改表

测试编号	6 号
测试类型	信息修改
前置条件	修改用户相关信息
预期结果	成功修改用户信息
实际结果	与预测结果一致

修改用户相关信息的测试结果如下图所示：

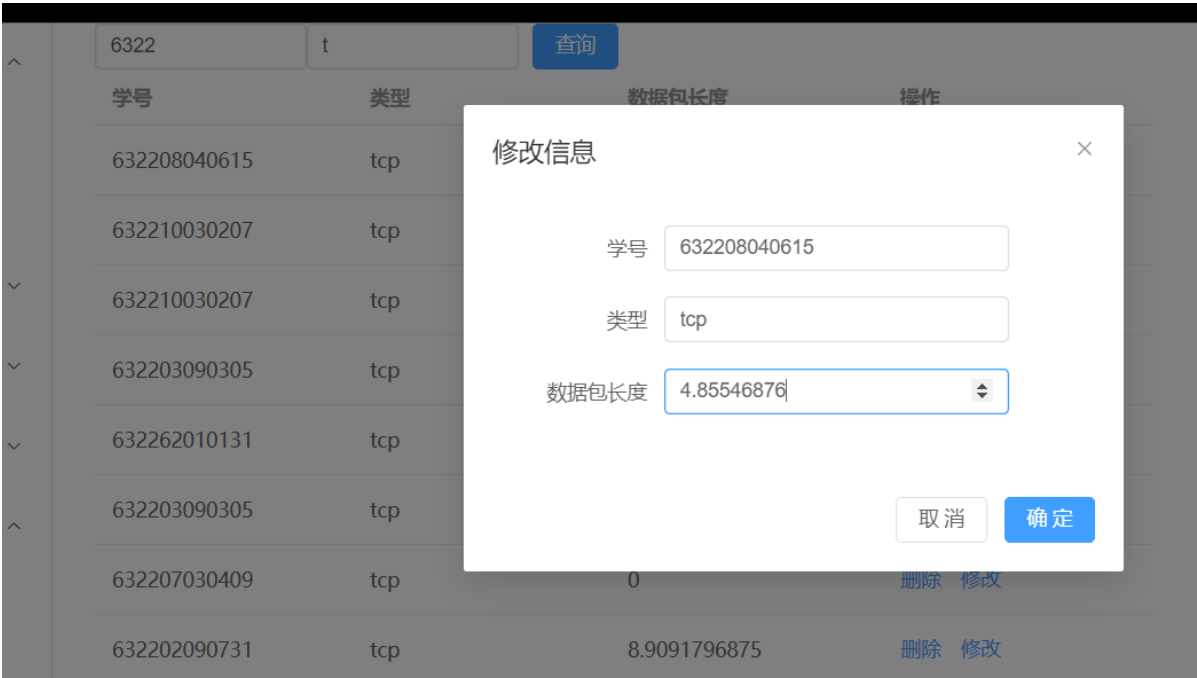


图 6.11 修改用户信息图



图 6.11 修改结果图

表 6.7 用户信息修改表

测试编号	7 号
测试类型	信息删除
前置条件	删除用户相关信息
预期结果	成功删除用户信息
实际结果	与预测结果一致

删除用户相关信息的测试结果如图 6.12 所示：



数据展示

删除成功

6322

t

查询

学号	类型	数据包长度	操作
632203090305	tcp	12.763671875	删除 修改
632262010131	tcp	16.0361328125	删除 修改
632203090305	tcp	0	删除 修改
632207030409	tcp	0	删除 修改
632202090731	tcp	8.9091796875	删除 修改
632205020117	tcp	6.107421875	删除 修改
632201150103	tcp	69.3310546875	删除 修改
632207060618	tcp	0	删除 修改
632208040203	tcp	0	删除 修改
632210030115	tcp	13.80078125	删除 修改

共 16565 条 10条/页 < 1 2 3 4 5 6 ... 1657 > 前往 1 页

图 6.12 删除用户信息结果图

6.4 本章小结

本章完成了对系统测试目的阐述，利用黑箱测试法对系统的重要功能进行了测试，得出的结果满足预期的要求，总体符合设计需求，测试通过。

## 第 7 章 总结与展望

随着互联网的飞速发展，互联网已经成为了我们生活密不可分的一个部分，伴随着我们对互联网的使用，相关数据也迎来爆发式的增长。在海量数据中，如何提取有价值的信息来提升我们的生活质量已成为研究的热点。

由于日常上网产生的网络数据具有高维度、多样性和复杂关联性，这使得数据结构变得复杂。选择合适的聚类方法来分析校园网用户行为数据，已成为研究的关键点。

本研究旨在评估不同聚类算法在校园用户行为研究中的表现，为规范学生上网行为和优化校园网络管理提供及时有效的建议。同时为应对不同属性的量纲问题本文根据数据的高时序性的特点完成了对数据的转换，为数据融合分析打下基础。最终通过可视化系统将相关结果展示出来，便于我们学校的相关管理人员直观地发现问题，及时作出应对。

本文主要完成了以下工作：

① 完成了对校园网用户日志数据的预处理，将一天分为 8 个计数点，把一周的用户日志数据中的网络使用时间与流量使用情况转换到对应的计数点，通过将不同属性的数据标准化消除了尺度差异，最后通过加权归一化的方法实现不同属性融合分析。同时计算用户对应的网络协议使用次数以及流量使用情况，对这两种属性完成上述的预处理过程。

② 对于预处理后的数据，比较 K-means++ 算法，DBSCAN 算法，以及自编码器的聚类效果，最终确定了 K-Means++ 算法来完成对用户的聚类分析。初步实现了对用户的上网时间喜好分析以及对于用户访问网站类型的偏好分析，帮助网络管理员更好地进行网络资源优化、安全管理、流量控制和用户支持，以提供更好的网络服务和用户体验。

③ 完成了系统的需求分析和设计，完成了用户访问协议可视化分析，各类用户的流量使用情况以及上网时间可视化分析。用户偏爱的网站类型分析。

本文仅仅是分析校园网学生用户群体网络行为的一些较为普遍的规律，还有很多地方可以进一步进行改善：

① 本研究主要分析了校园网中学生用户的网络行为，但尚未涵盖教职工群体。教职工的网络使用模式同样值得研究，因为它们可以揭示工作时间内的在岗工作状态，为人事管理部门提供有价值的数据库支持。

② 可以针对学生的成绩与网络的依赖程度间的关联进行分析，得到成绩与网络使用情况的关系。

③ 可以使用分布式的操作系统来对系统进一步改善，实现对大数据的实时处理。

## 致 谢

## 参 考 文 献

- [1] 中国互联网络信息中心发布第 52 次《中国互联网络发展状况统计报告》[J]. 国家图书馆学刊, 2023, 32(05):13.
- [2] 马仕玉. 聚类算法及其在校园网用户行为分析中的应用[D]. 重庆交通大学, 2016.
- [3] 贺雯静. 校园网用户行为分析系统的设计与实现[D]. 西北大学, 2021.
- [4] Zeinab L, Amir HR, Ala M. Application of data mining methods for link prediction in social networks[J]. Social Network Analysis and Mining, 2013, 3(2):143–150.
- [5] Xhafa F, Santi Caballé, Barolli L, et al. Using Bi-clustering Algorithm for Analyzing Online Users Activity in a Virtual Campus[C]//International Conference on Intelligent Networking & Collaborative Systems. IEEE, 2011.
- [6] Malvika Singh, B.M. Mehtre, S. Sangeetha. User Behavior Profiling using Ensemble Approach for Insider Threat Detection[C]//2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis: IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), 22-24 Jan. 2019, Hyderabad, India. Institute of Electrical and Electronics Engineers, 2019:1-8.
- [7] Soledad D, Federico M, San C J, et al. Analysis of Students' Behavior Through User Clustering in Online Learning Settings, Based on Self Organizing Maps Neural Networks[J]. IEEE ACCESS, 2021, 9, 132592-132608.
- [8] Al-Mashhour, Alaa. Machine-Learning-based User Behavior Classification for Improving Security Awareness Provision[J]. International Journal of Advanced Computer Science and Applications, 2023, 14(8):166-178.
- [9] 宋坤. 校园网用户行为信息的预处理与聚类方法研究[D]. 重庆交通大学, 2017.
- [10] 朱峦. 基于聚类分析的校园网用户行为分析系统的设计与实现[D]. 湖北工业大学, 2017.
- [11] 祁家祯. 基于大数据的用户分析系统的设计与实现[D]. 北京交通大学, 2021.
- [12] 陈新泉, 周灵晶, 刘耀中. 聚类算法研究综述[J]. 集成技术, 2017, 6(03):41-49.
- [13] 钱杨. 对计算机软件测试技术的几点探讨[J]. 电子测试, 2021(3):2.
- [14] 王玲, 薄列峰, 焦李成. 密度敏感的半监督谱聚类[J]. 软件学报, 2007, (10):2412-2422.
- [15] 唐恺. 面向企业竞争情报的文本聚类技术的研究与应用[D]. 西安电子科技大学, 2013.

- [16] 骆盈盈, 陈川, 毛云芳. 基于传感器网络的 K-均值聚类算法研究[J]. 计算机工程与设计, 2007(06):1349-1351.
- [17] 李学锋. 校园网建设项目可持续性评价分析[J]. 电脑迷, 2018(09):181.
- [18] 高璐瑶. 安全苛求软件的自动化测试技术研究[D]. 浙江大学, 2013.
- [19] 张超永, 姬成群, 时晓宁. 对白盒测试中动态测试技术研究[J]. 电脑编程技巧与维护, 2015(15):25-26.