陈万志,赵宇璇. 智慧校园隐式用户行为的数据挖掘方法[J].辽宁工程技术大学学报(自然科学版),2020,39(5):434-439. doi:10.11956/j.issn.1008-0562.2020.05.009
CHEN Wanzhi, ZHAO Yuxuan. Data mining method for implicit user behavior in smart campus[J].Journal of Liaoning Technical University(Natural Science), 2020,39(5):434-439. doi:10.11956/j.issn.1008-0562.2020.05.009

智慧校园隐式用户行为的数据挖掘方法

陈万志，赵宇璇
（辽宁工程技术大学 电子与信息工程学院，辽宁 葫芦岛 125105）

摘　要：为解决通过智慧校园隐式用户行为信息构建学生画像的数据挖掘问题，提出一种聚类分析与神经网络相结合的方法.以学生账号为关键字分析提取用户行为数据中的上网总时长、总流量和上下线时间点的映射扩展分别作为密度聚类和谱聚类的特征，通过数据预处理得到用户的上网行为标签；采用遗传算法优化的神经网络实现用户个性化兴趣偏好的判别预测；通过词云图的形式实现学生个人画像的可视化.研究结果表明：该方法的挖掘效果优于传统的神经网络.研究结论进一步完善了学生画像的构建和分析汇总，能够有效地支撑学生管理和服务.

关键词：智慧校园；用户画像；聚类分析；遗传算法；神经网络；词云图
中国分类号：TP 391.1　　**文献标志码**：A　　**文章编号**：1008-0562(2020)05-0434-06

Data mining method for implicit user behavior in smart campus

CHEN Wanzhi, ZHAO Yuxuan
(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: In order to solve the problem of data mining for constructing student portraits by using the implicit user behavior information on intelligent campus, a method combining clustering analysis and neural network is proposed. Firstly, the total online surfing time, total traffic, and the mapping extension of the up and down time points in the user behavior data which are analyzed and extracted by using student account as keyword are used as characteristics of density clustering and spectral clustering respectively, then the user's online behavior label is obtained through data preprocessing. Secondly, the neural network optimized by genetic algorithm is used to identify and predict users' personalized interest preference. Finally, the visualization of students' personal portrait is realized through the form of word cloud map. The research results show that the mining effect of this method beats the traditional neural network. The research conclusion further improves the construction and analysis summary of student portrait, which can effectively support student management and service.

Key words: smart campus; user portraits; clustering analysis; genetic algorithms; neural network; word cloud

0　引言

目前国内外高校对大数据研究与应用的重要性已达成共识，教育数据中心建设、区域数据共享和应用已成为主要的实践模式.高校正逐步从数字校园转向智慧校园建设，教育大数据成为研究热点，学生隐式行为数据作为重要的组成部分，能够全面、系统地反映学生的行为规律和特征，对高校教育管理、学生自身发展具有重要意义.大数据时代，如何挖掘用户的特征和行为规律，特别是基于智慧校园的隐式用户行为分析挖掘，不仅便于学生认知自身偏好，也为管理者掌握学生课堂出勤、消费情况和夜不归宿等实际需求，甚至为涉及学生利益的奖贷困补入党推免等工作提供可参考的技术性依据.

学者们针对用户行为数据挖掘的研究热点提出一系列算法模型.文献[1]～文献[4]基于学校一卡通数据利用 *K*-means 算法对学生聚类划分，或通过决策树预测学生行为指标进行等级预警，能够有效

收稿日期：2019-10-15
基金项目：辽宁省教育厅服务地方类项目（LJ2017FAL009）；辽宁工程技术大学博士启动基金（2015-1147）
作者简介：陈万志（1977-），男，辽宁 阜新人，博士，副教授，主要从事人工智能与智能信息处理、大数据与云计算、网络与信息安全、计算机控制与嵌入式软件、物联网工程等方面的研究.

区分不同行为特征用户.但是依据单属性进行行为分析.文献[3]以 IP 地址进行聚类，并进一步对分类的用户群进行分析，但缺乏数据的实时性.文献[5]根据用户画像模型挖掘银行客户的产品、理财、风险等偏好，然后按等级对用户进行个性化推荐.用户画像已运用在学生生活学习、商品推荐以及用户行为轨迹等方面，对深刻认识个人的作息规律及潜在习惯有一定的指导意义[5-8].

综上所述，本文提出一种智慧校园隐式用户行为的数据挖掘的模型，以学生的上网日志明细数据为基础，充分考虑学生作息及上网的实时性，通过密度聚类和谱聚类进行特征提取，再利用遗传算法优化神经网络的权值阈值的 GA-BP（Genetic algorithms-Back propagation）方法提高用户个性化兴趣偏好的判别预测，最后将学生个人静态和动态信息[9-10]利用词云图的形式实现可视化综合展示.

1　总体框架

智慧校园隐式用户行为的数据挖掘总框架见图 1，即通过收集学生校园卡系统、门禁系统、学生管理系统和教务系统等服务数据，无线网络、互联网和 URL 等日志数据，各类门户网站、新闻、微博、社交、论坛等外部互联网访问数据[11]，利用所提出的算法模型构建学生用户画像，对学校的教学、日常管理、学生指导等方面提供科学有力的依据[12].

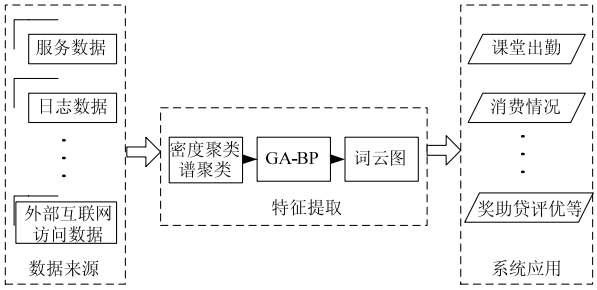

图 1　系统总体框架

Fig.1　overall framework of the system

2　算法模型

2.1　算法流程

算法首先提取学生用户相关数据进行预处理，其次用密度聚类和谱聚类算法进行特征提取，再次利用遗传算法优化的神经网络进行个性化兴趣偏好的判别预测，最后以词云图的形式实现学生用户个人画像的可视化.算法总体流程见图 2.

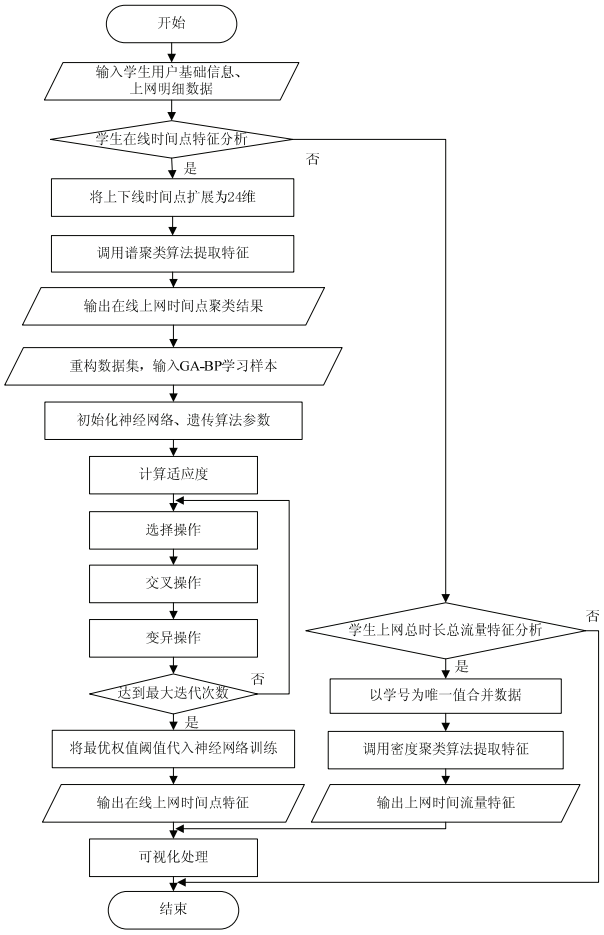

图 2　算法总体流程

Fig.2　overall flow chart of algorithm

2.2　密度聚类 DBSCAN

为保证智慧校园的隐式用户行为分析挖掘的可靠性和实时性，以学生账号为关键字，提取用户行为数据中的上网总时长、总流量作为密度聚类特征，通过调用 Pycharm numpy 库中的 DBSCAN 得到用户的上网行为标签.

2.3　谱聚类 Spectral Clustering

谱聚类具备能够有效的避免陷入局部最优，数据适应能力较强、计算量较小和聚类效果较好的优点.由于在智慧校园隐式用户行为的数据挖掘中所提取的学生上网日志明细数据的特征具有一定的局限性，故将上下线时间点映射扩展为 24 维，以此作为谱聚类算法的输入，其中每个维度代表 24 小时中的一个时间点，输出则为在线时间点特征，代表用户的上网行为标签.在线时间点特征提取流程见图 3.

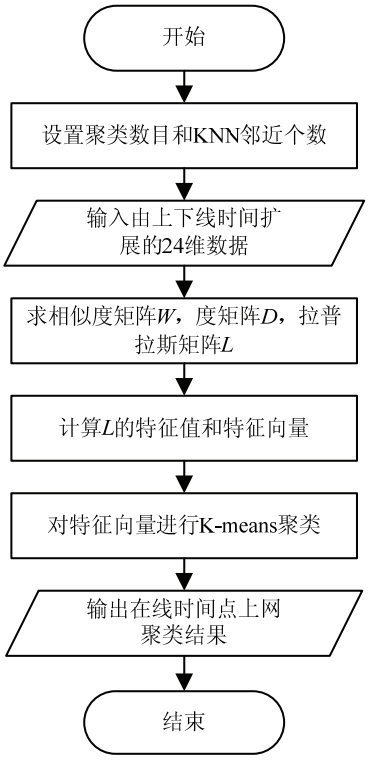

图 3 在线时间点特征提取流程

Fig.3 flow chart of online feature extraction

定义 1 欧式距离.用来度量样本之间的相似性.任意两个学生 i、j 的在线时间点扩展 24 维向量后的相似度为

$$d(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2},\quad(1)$$

式中，x_i、x_j 分别为学生 i、j 的在线时间点特征向量，$n\in[1,24]$.

定义 2 KNN.用来构造相似度矩阵 $\boldsymbol{W}$.

通过遍历所有学生在线时间点特征向量扩展构造向量矩阵 $\boldsymbol{W}'$

$$W'_{ij}=W'_{ji}=\begin{cases}0 & i=j\\ d(x_i,x_j) & i\neq j\end{cases}.\quad(2)$$

对 $\boldsymbol{W}'$按行进行升序排列，根据式（3）重构矩阵 $\boldsymbol{W}''$

$$W''_{ij}=\begin{cases}1 & j\text{为}i\text{的}k\text{邻近点}\\ 0 & \text{其他}\end{cases}.\quad(3)$$

$\boldsymbol{W}''$是非对称矩阵，将其与对应的转置矩阵相加再求平均得到对称矩阵 $\boldsymbol{W}$.

定义 3 度矩阵 $\boldsymbol{D}$.由 $\boldsymbol{W}$ 的每一行元素之和 d_i 构成度矩阵 $\boldsymbol{D}$

$$d_i=\sum_{i,j=1}^{n}W_{ij}\quad 1\leqslant n\leqslant 24.\quad(4)$$

定义 4 拉普拉斯矩阵 $\boldsymbol{L}$

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}.\quad(5)$$

2.4 GA-BP 算法

为进一步提高用户个性化兴趣偏好的判别预测，采用遗传算法优化神经网络，克服迭代因陷入局部极小陷阱形成死循环的缺点[13]，进而获得最优权值和阈值.遗传算法优化神经网络算法流程见图 4[14].

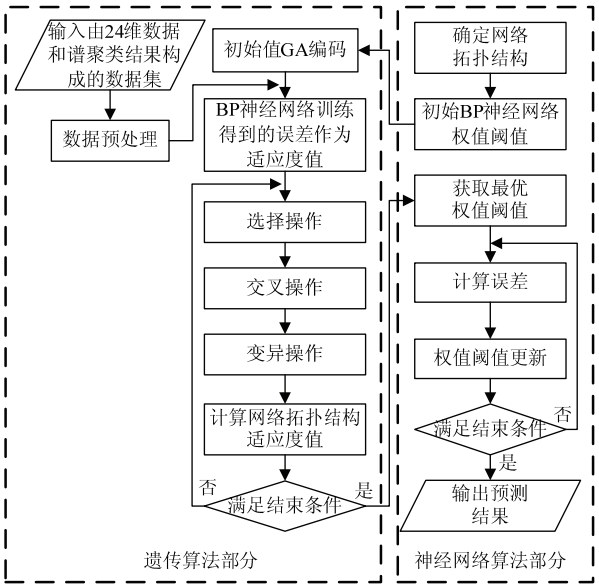

图 4 遗传算法优化神经网络算法流程

Fig.4 flow chart of genetic algorithm optimization neural network algorithm

2.5 词云图可视化展示

在智慧校园隐式用户行为的数据挖掘预测结果采用词云图的形式进行用户信息的可视化综合展示.词云图可视化实现的步骤为：

（1）获取词的内容，即学生个人信息；

（2）获取内容展示区域的遮罩图；

（3）调用 wordcloud 库显示词云.

3 实验

3.1 实验环境

实验测试环境为 Intel i5-4210u 2.40 Hz，操作系统为 Windows10 专业版，内存为 8 G，软件为 Pycharm2018.

3.2 数据集

（1）数据采集

实验数据为辽宁工程技术大学校园网上网明细日志部分数据，包含账号、总流量、总时长、入流量、出流量、上线时间、下线时间、MAC 地址和 IP 地址

等 16 个维度.考虑学生作息时间及上网的特性，选取 2019 年 5 月 14 日 0 点到 19 点上线的共 12 444 条上网日志明细作为测试数据集，见表 1.

表 1　部分采集数据

Tab.1　partial data collection

账号	总流量/M	总时长/(h-m-s)	…	上线时间	下线时间	…
1720010228	1.62×10^3	11-27-30	…	2019-05-14 07:29:59	2019-05-14 18:57:29	…
1629020102	493.38	10-39-1	…	2019-05-14 08:18:32	2019-05-14 18:57:33	…
1821940220	17.22	1-10-39	…	2019-05-14 17:46:55	2019-05-14 18:57:34	…
1810210413	90.41	4-23-16	…	2019-05-14 14:34:35	2019-05-14 18:57:51	…
1720011021	784.98	2-54-41	…	2019-05-14 16:03:14	2019-05-14 18:57:55	…
1610200108	2.19×10^3	10-32-44	…	2019-05-14 08:25:13	2019-05-14 18:57:57	…
1703030116	784.67	4-33-31	…	2019-05-14 14:24:33	2019-05-148:58:04	…

（2）数据处理

智慧校园隐式用户行为的数据来源复杂，冗余度高，在进行潜在的信息挖掘之前必须进行数据清洗、单位化和整合等数据预处理.

测试实验对象主要针对辽宁工程技术大学葫芦岛校区的本科生，故首先去除非挖掘对象的日志数据及挖掘对象的日志数据中的 MAC 地址、下线原因、产品 ID 等属性数据.其次依据数据集中维度属性的特征特点，以学生账号为关键字分别构建数值型和属性划分型两个子集，即将同一账号的总流量、总时长属性数据分别进行累计合并量化；将上下线时间映射扩展为 24 维，若上线时间的分和秒有一个非零，则起始时间点取上线时间点的小时值加一，否则取当前小时值；结束时间点取下线时间对应的小时值，如上线时间 5:30:20，下线时间 9:20:15，则在线时间点选取为[6,9].初始化 24 维在线时间点向量为 [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]，设置在线时间点为 1，离线时间点为 0，则结果为[0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0].

3.3　结果分析

将数据处理得到的数值型数据子集作为密度聚类的特征，因无真实标签，故采用内部指标轮廓系数 *sc* 评估聚类效果，即

$$sc=\frac{b-a}{\max(a,b)}. \tag{6}$$

同一类别中个样本点与其他样本的平均距离记为 *a*，个样本点与距离最近不同类别中的样本的平均距离记为 *b*.*sc* 取值范围为[-1,1]，越接近 1 则聚类效果越好.

选取上网总时长和上网总流量分别作为密度聚类的特征进行单属性聚类分析，结果见图 5，其中 *k* 代表聚类个数，k_1 和 sc_1 是上网总流量聚类结果，k_2 和 sc_2 是上网总时长聚类结果.可知随着邻域半径 *eps* 的增加 *k* 先下降后稳定，*sc* 处于均衡变化.将上网总时长和总流量综合进行多属性密度聚类，结果见图 6，其中 *eps*=0.4 时，聚类效果最好，对应的 *sc* 值为 0.774，*k* 值为 3，基本符合智慧校园隐式用户上网行为实际情况，后者的聚类效果优于前者.

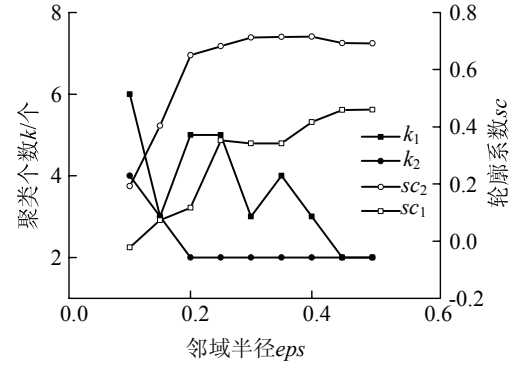

图 5　单属性密度聚类

Fig.5　single attribute DBSCAN clustering

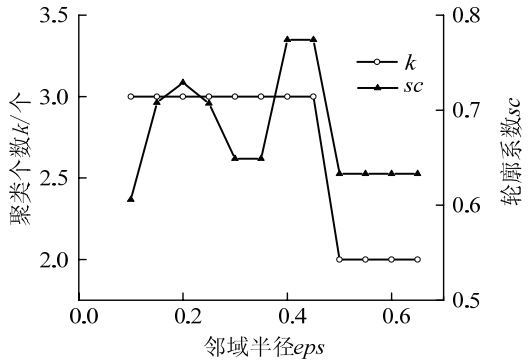

图 6　多属性密度聚类

Fig.6　multi-attribute DBSCAN clustering

将数据处理得到的属性划分型数据子集作为谱聚类的特征，图 7 为本方法 KNN 邻近点个数选

取对比结果，当聚类个数 k 一定时，随着邻近点个数的增加 sc 先上升后下降；当邻近点个数确定时，k 越小聚类效果越好.k 值为 3，邻近点数为 10 时，聚类效果最好，基本符合用户上网行为的实际情况. 图 8 为属性个数聚类对比结果，其中单属性 1 聚类为仅选择上线时间、单属性 2 聚类为仅选择下线时间，多属性聚类为上、下线时间综合.可知单属性聚类不具备代表性，故本文方法采用多属性聚类方式.

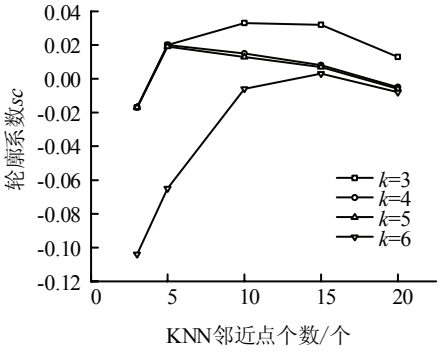

图 7 邻近点个数聚类对比

Fig. 7 comparison of number of adjacent points clustering

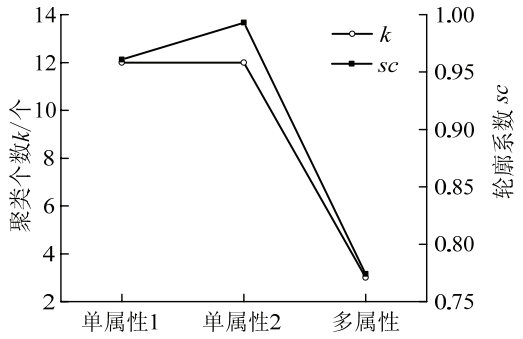

图 8 属性个数聚类对比

Fig.8 comparison of attribute number clustering

在 Windows10 操作系统和 Python3.7 的环境下，基于由 24 维数据和谱聚类结果构成的数据集，分别采用遗传算法优化神经网络算法（GA-BP）、神经网络算法（BP）、分类和回归树算法（CART）、逻辑回归算法（LR）、支持向量机算法（SVM）和高斯朴素贝叶斯算法（NB）进行判别预测，采用 k 折交叉验证结果的平均值作为模型的性能评估指标[15]，同一算法 k 由 5 变为 10，平均值略微下降，当 k 值一定时，BP 神经网络算法优于其他四个机器学习算法，且 GA-BP 算法优于 BP 神经网络算法. 各种算法 k 折交叉验证准确度对比见图 9，可知神经网络在智慧校园隐式用户个性化兴趣偏好的判别预测方面优于机器学习算法，且本文采用优化的 GA-BP 算法对智慧校园隐式用户个性化兴趣偏好的判别预测效果最优.

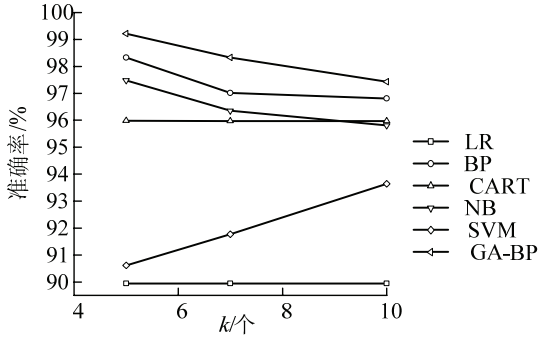

图 9 各种算法 k 折交叉验证准确度对比

Fig.9 comparison of accuracy of k folding and cross validation of various algorithms

根据由智慧校园隐式用户行为信息构建学生画像的需求，本文采用用户个性化兴趣偏好的判别预测准确率为

$$\text{准确率}=\frac{\text{正确预测的学生数}}{\text{所有参与预测的学生总数}}. \qquad (7)$$

利用遗传算法优化神经网络，经过反复测试，选取种群数为 60，交叉率为 0.8，突变率为 0.1，遗传算法优化后的神经网络与传统神经网络完成学生用户个性化兴趣偏好的判别预测准确率对比见表 2，可知优化后的神经网络判别预测准确率高.

表 2 准确率对比

Tab.2 comparison of accuracy

算法	准确率/%
BP	0.983 235
GA-BP	0.992 187

3.4 可视化展示

数据挖掘预测结果采用词云图的形式进行用户信息的可视化综合展示，满足大数据分析平台的人机交互需要.以某同学为例，可视化展示的静态信息包括学生姓名、学院、专业、生源地等基本信息，动态信息包括聚类结果标签化，即由网络日志数据中数值型特征判别赋予“网瘾少年”的标签，属性划分型特征判别赋予“长时间在线作息可能不合理”的标签[10]，以及应用流量详细信息汇总饼图，学生个人用户画像可视化见图 10.

4 结论

针对解决智慧校园隐式用户行为信息构建学生画像的数据挖掘问题，提出一种聚类分析与神经网络相结合的算法模型，根据具体工作流程可知该方法的优点如下：

（1）挖掘效果优于传统的神经网络；

（2）所构建的学生用户画像和分析汇总更利于学生认知自身偏好，教师开展学生管理工作，具有一定的实际意义；

（3）本文的方法同样适用于社会人群的研究.

本文对智慧校园隐式用户行为的研究还有待进一步深入，下一步将会从消费记录、位置信息等更多维度进行挖掘.

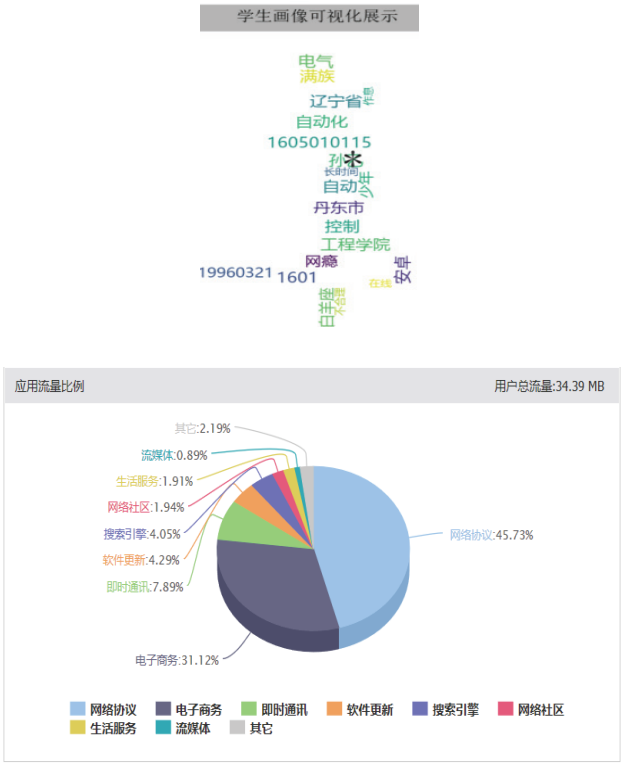

图 10　学生个人用户画像可视化

Fig.10　visualization of student personal user portrait

参考文献（References）：

[1] 梁柱.基于校园大数据的学生行为分析与预测方法研究[D].西安:西安理工大学,2017.

[2] 董潇潇,胡延,陈彦萍.基于校园数据的大学生行为画像研究与分析[J].计算机与数字工程,2018,46(6):1 200-1 204,1 262.
DONG Xiaoxiao,HU Yan,CHEN Yanping.Research and analysis of behavior portrait of college students based on campus data[J].Computer and Digital Engineering,2018,46(6):1 200-1 204,1 262.

[3] 李游.基于校园网的用户行为分析研究[D].昆明:云南大学,2013.

[4] 黄刚,刘蓉,刘合富,等.基于校园一卡通数据的人群画像分析[J].计算机与数字工程,2018,46(9):1 881-1 886.
HUANG Gang,LIU Rong,LIU Hefu,et al.Crowd portrait analysis based on campus one-card data[J].Computer and Digital Engineering,2018,46(9):1 881-1 886.

[5] 孙晔,杨照东,陈德华,等.大数据用户画像技术在商业银行的应用[J].数字通信世界,2016(9):86-88.
SUN Ye,YANG Zhaodong,CHEN Dehua,et al.Application of big data user portrait technology in commercial Banks[J].Digital Communications World,2016(9):86-88.

[6] 王庆福.贝叶斯网络在用户兴趣模型构建中的研究[J].无线互联科技,2016(12):101-102.
WANG Qingfu.Research on the construction of bayesian network in user interest model[J].Wireless Interconnection Technology,2016(12):101-102.

[7] 黄文彬,徐山川,吴家辉,等.移动用户画像构建研究[J].现代情报,2016,36(10):54-61.
HUANG Wenbin,XU Shanchuan,WU Jiahui,et al.Research on the construction of mobile user portrait[J].Modern Intelligence,2016,36(10):54-61.

[8] 王菊艳.基于 WEB 日志的用户画像及可视化分析[D].西安:西安理工大学,2019.

[9] 万家山,陈蕾,吴锦华,等.基于 KD-Tree 聚类的社交用户画像建模[J].计算机科学,2019,46(S1):442-445,467.
WAN Jiashan,CHEN Lei,WU Jinhua,et al.Social user portrait modeling based on KD-Tree clustering[J].Computer Science,2019,46(S1):442-445,467.

[10] 宋美琦,陈烨,张瑞.用户画像研究述评[J].情报科学,2019,37(4):171-177.
SONG Meiqi,CHEN Ye,ZHANG Rui.Review on user portrait research [J].Intelligence Science,2019,37(4):171-177.

[11] SU Hui G,CHENG Jie B,QUAN W.Hadoop-based college student behavior warning decision system[C]//2018 IEEE 3rd International Conference on Big Data Analysis.Shanghai:Institute of Electrical and Electronics Engineering, Inc.,2018:217-221.

[12] 勾志竟,任建玲,徐梅,等.基于 Hadoop 的 GA-BP 算法在降水预测中的应用[J].计算机系统应用,2019,28(9):140-146.
GOU Zhijing,REN Jianling,XU Mei,et al.Application of ga-bp algorithm based on Hadoop in precipitation prediction[J].Application of Computer System,2019,28(9):140-146.

[13] 谢梦蝶,秦江涛.遗传算法优化 BP 神经网络预测股指研究[J].软件导刊,2019,18(4):41-45.
XIE Mengdie,QIN Jiangtao.Study on BP neural network prediction index by genetic algorithm optimization[J].Software Guide,2019,18(4):41-45.

[14] 康海燕,杨悦,于爱民.面向用户的电商平台刷单行为智能检测方法[J].计算机应用,2018,38(2):596-601.
KANG Haiyan,YANG Yue,YU Aimin.Intelligent detection method of brushing behavior on user-oriented e-commerce platform[J].Computer Application,2018,38(2):596-601.