文章编号：2095-2163(2022)10-0200-03 中图分类号：TP393 文献标志码：A

基于模糊 K 均值聚类的高校网络用户行为分析

于莉佳[1]，汪 涛[2]
（1 中国联合网络通信有限公司哈尔滨软件研究院，哈尔滨 150040；2 哈尔滨商业大学，哈尔滨 150028）

摘 要：随着网络科技的迅猛发展，互联网用户的规模正在以指数的速度不断增长。高校网络用户的规模也随着互联网兴起而出现大规模增长。对高校网络用户的上网行为进行分析，能够更好地掌握在校学生的动态，为学校制定科学、高效的互联网管理方式奠定了更加客观的数据基础。本文首先将高校网络用户上网行为进行分类，然后通过模糊 K 均值聚类算法对学生的上网行为进行分类。实践表明，通过对某高校的学生上网行为展开分析，为该校的互联网管理和学生的精细化管理提供了有利的数据支撑。

关键词：模糊 K 均值聚类算法；网络用户行为；数据挖掘

Analysis of university network users behavior based on fuzzy K-means clustering

YU Lijia[1], WANG Tao[2]
(1 Harbin Software Research Institute of China Unicom, Harbin 150040, China;
2 Harbin University of Commerce, Harbin 150028, China)

【Abstract】 With the rapid development of network technology, the scale of Internet users is growing exponentially. Consequently, the scale of network users in colleges and universities has also grown massively with the rise of the Internet. By analyzing the online behavior of network users in colleges and universities, the dynamics of students in schools coud be better grasped, and more objective data foundation could be established for schools to formulate scientific and efficient Internet management methods. This paper first classifies the online behaviors of college network users, and then uses the fuzzy K-means clustering algorithm to classify the online behaviors of students. Practice shows that analyzing the online behavior of students in an university could provide favorable data support for the school's Internet management and refined management of students.

【Key words】 fuzzy K-means clustering algorithm; network users behavior; data mining

0 引 言

随着互联网技术的高速发展，人们的生活已经与互联网息息相关。在互联网迅速普及的大背景下，互联网用户的行为分析现已成为洞察用户偏好、用户能力的一个有利手段[1]。用户网络行为的分析给网络平台提供了更加多元的选择，但与此同时也为网络平台提出了更加严苛的技术要求和规范。网络用户的互联网行为被数据平台监控，平台则通过数据分析来了解用户的意图，从而促进网络生态环境的良性发展[2-4]。

目前，认证计费、流量监控等应用服务器已然广泛应用在各大高校的网络管理中。这些应用服务器在为高校提供管理便利的同时，却还会产生大量的日志数据，这些日志数据通常存储于后台数据库当中。分析可知，日志数据包含大量的用户在互联网上的行为数据。如果能对日志中行为数据进行科学、高效分析，并且对数据隐藏的深层次的规律加以利用，将大大地提升高校的网络管理效率，为高校网络管理构建有效支撑，且为其决策科学化、管理精细化提供有益帮助。本文以一具体高校为实例，对用户上网行为数据进行聚类分析，挖掘内在规律，助力高校决策顺利实施。

1 网络用户行为分类

针对高校网络用户行为可以归纳为不同的类别[5]。分类类别与研究方法和研究目的密切相关。比如，根据网络用户网络行为流量异常情况，可以将用户分为流量正常用户和流量异常用户，流量异常用户的计算机通常已被蠕虫、木马等病毒感染。针对用户的网络行为攻击情况，可以分为善良网络用户和恶意网络用户。由此即可展开善良网络行为推演，分析其访问偏好、常用的网络访问模式，从而更好地引导高校网络用户正确使用互联网，指导运营商为高校用户提供更为优质的、更有针对性的互联网接入服务[6-7]。基于本文的研究问题及目的，则

作者简介：于莉佳(1980-)，女，硕士，主要研究方向：数据挖掘、图像处理、模式识别；汪 涛(1987-)，男，博士，主要研究方向：大数据、模式识别、医学图像处理。

收稿日期：2022-03-29

将高校的学生的网络行为分为 4 类，即：信息获取类行为、知识获取类行为、休闲娱乐类行为、电子商务类行为。对此拟做探讨阐述如下。

(1)信息获取类行为。是用户通过使用超文本协议等从互联网上查询并获取自身所需要的信息。互联网将数以万计、不同格式、零散的信息进行整合。信息获取类行为最大的特点是通过搜索引擎来收集互联网上的资源。

(2)知识获取类行为。高校学生知识获取类行为主要是指从不同的课程平台获取知识的行为。目前线上教学已经成为高校教学活动中的一个重要的组成部分。学堂在线、网易云课堂、果壳 MOOC 学院等一批慕课平台陆续出现，吸引了大量的在线学员。高校学生从在线平台获取知识的行为日趋频繁，该知识获取行为也已成为线下课堂教学的重要补充。

(3)休闲娱乐类行为。休闲娱乐类行为把互联网视为一个开放的娱乐场所，可提供多种服务用于消遣。比如，互联网可以为用户提供玩游戏、看电影、看小说、听音乐等多种服务。通常情况下，休闲娱乐行为是电子商务行为的子集。但是考虑到休闲娱乐类行为占用了互联网用户网络行为的很大比例，因此就将休闲娱乐类行为作为高校学生网络行为的一个单独分类。

(4)电子商务类行为。电子商务类行为把互联网当成一个开放的交易场所，可为互联网用户提供信息获取功能、沟通交流功能。电子商务公司通过在互联网上建立虚拟交易平台，为网络用户提供交易场所，如亚马逊、淘宝、易趣网、京东等。目前，高校学生消费方式已经从线下购买转移到网上购物，并将网络上的虚拟交易平台作为购物的主要方式。因此，本文将电子商务类行为作为高校学生上网行为的一个重要分类。

2　算法研发

FCM 聚类算法由 Dunn 等人提出，Li 等人对该算法进行了改进。FCM 算法中不同样本点对聚类中心有一个在[0,1]范围内的隶属度，根据隶属度的大小对样本点的类别进行划分。

定义网络用户行为样本 $I=[I_1,I_2,\cdots,I_N]$，I^s 表示第 s 类别网络行为周时长(单位：小时)比例。这里，$s\in[1,2,\cdots,N]$。根据网络用户行为样本定义可知，网络用户 I_n 的 s 行为所占的周时长比例用 I_n^s 表示。

由 FCM 聚类算法选择的相似度函数可知，将 I_n 聚类为 K 个类别，其中聚类中心称为 V_k，$k=1,2,\cdots,K$。定义 I_n 与类别的隶属度，即模糊划分矩阵，为 $\boldsymbol{U}=[u_{ik}]_{c\times n}$，隶属度应满足下列约束条件：

$$u_{ik}\in[0,1],\ 1\leqslant i\leqslant n,\ 1\leqslant k\leqslant c \tag{1}$$

$$\sum_{k=1}^{c}u_{ik}=1,\ 1\leqslant i\leqslant n \tag{2}$$

其中，u_{ik} 表示样本 I_n 与类别 k 的隶属度。

研究中又给出了 FCM 算法的目标函数定义如下所示：

$$J(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ik})^2d^2(X_i,v_k) \tag{3}$$

其中，d 表示欧氏距离函数。

进一步地，定义聚类中心迭代更新函数如式(4)所示：

$$v_k=\frac{\sum_{i=1}^{n}(X_iu_{ik}^2)}{\sum_{i=1}^{n}u_{ik}^2} \tag{4}$$

这里，将对基于 FCM 聚类的网络用户上网行为聚类算法的步骤流程做全面表述如下。

算法 1　基于 FCM 聚类的网络用户上网行为聚类算法

输入　网络用户行为样本 I

输出　网络用户行为聚类结果

步骤 1　给定待聚类样本聚类数目 K 及相关参数，本文中聚类数目 $K=4$。

步骤 2　初始化隶属度矩阵 $\boldsymbol{U}$ 及 K 个聚类中心。

步骤 3　计算待聚类样本 I 与聚类中心距离矩阵，并更新隶属度矩阵 $\boldsymbol{U}$。

步骤 4　对目标函数 J 进行计算，如果小于给定阈值 δ，则进入步骤 5，否则返回步骤 2 继续迭代。

步骤 5　计算全部样本 I 的最近距离聚类中心，更新样本 I_n 的类别为最近距离聚类中心类别号。

3　结果分析

目前仍然无法量化分析高校学生的网络行为，上网行为评价的各类信息仅仅停留在主观评价方面。高校赠予学生用于访问学习资源的免费流量存在较难评判的客观性。部分学生存在虚假申请免费流量问题，导致有限的资源无法分配给有需求的学生，造成网络资源的严重浪费。这里通过学生网络行为进行分析，将学生分为 4 种类型：学习型、学习

游戏型、消费娱乐型、游戏型。

本实验基于某高校互联网用户行为数据构建的数据集,从中随机选择 1 000 个用户,采集 3 月份第三周的网络日志文件,统计这 1 000 个学生用户的信息获取类行为、知识获取类行为、休闲娱乐行为、电子商务类行为。利用 K 均值模糊聚类算法将产生这些网络行为的用户分为学习型、学习游戏型、消费娱乐型以及游戏型,统计聚类结果见表 1。

表 1 网络用户行为聚类结果

Tab. 1 Clustering results of online users behavior %

	信息获取类	知识获取类	休闲娱乐类	电子商务类	人数比例
学习型	20	65	5	10	65
学习游戏型	10	34	37	19	21
消费娱乐型	5	22	42	31	13
游戏型	5	7	78	10	1

由表 1 中该高校学生的上网行为可以看出,学习型和学习游戏型占比较高,总计占比 86%。学习型用户网络行为的时间主要表现为信息获取和使用网络学习资源。学习游戏型用户网络行为主要以使用网络学习资源和游戏为主。消费娱乐类行为的用户主要的网络行为集中体现在休闲娱乐和电子商务类行为这 2 个方面,而这 2 类行为占比达到了 73%。最后一类用户是游戏型用户,这类用户的主要网络行为是休闲娱乐类,大多表现为网络游戏行为。

通过本高校用户网络行为聚类结果分析可知,该校的大部分学生的网络行为和学生有关,比例达 86%。本文针对这部分学生给予一定的免费上网时长,这样可以激励其更加倾向于自主从事网络学习行为来提升自己。即使是消费娱乐类学生的知识获取行为占比也达到了单周上网总时长的 22%。这也说明消费娱乐型学生在消费娱乐的同时,仍有一定的学时花费用于学习。最后一类游戏型学生,在休闲娱乐类的行为达到了单周上网时长的 78%,就说明这些学生大部分上网选择的都是休闲娱乐类行为。对这些学生的上网数据通过进一步分析发现,几乎所有的网络行为都是网络游戏。对于这部分学生需要引起学院的重视,并由辅导员给予这些学生重点关注。综上可知,这样一来,就可以更加精确地掌握学生的动态以及生活学习情况。

4 结束语

本文针对某高校学生上网行为的数据进行了挖掘、分析与探索。首先对高校学生上网行为分为信息获取类行为、知识获取类行为、休闲娱乐类行为、电子商务类行为四大类。然后利用 K 均值模糊聚类算法对学生的上网数据进行聚类分析。根据聚类结果挖掘、分析出数据所蕴含的更深层次的信息。借助本文研究成果,高校可以更加全面、客观地制定出相关互联网管理策略,同时也可以准确可靠地掌握学生状态,更有针对性地关注学生的健康成长。

参考文献

[1] 祁薇熹,李彬.多目标演化算法的进展研究[J]. 计算机与数字工程,2008,36(05):16-18,26.
[2] 丁青,周留根,朱爱兵,等. 基于 K-Means 聚类算法的校园网用户行为分析研究[J]. 微计算机应用,2010,31(06)74-80.
[3] 刘纯平. 基于 Kohonen 神经网络聚类方法在遥感分类中的比较[J]. 计算机仿真,2006,26(07):1744-1746,1750.
[4] 数据堂中心. 网络用户行为日志集[R/OL]. [2020]. http://www.datatang.com/data/43910.
[5] 孙家广,杨长青. 计算机图形学[M]. 北京:清华大学出版社,1995.
[6] 李大伟. 数据挖掘在用户行为分析中的研究与应用[D]. 北京:北京邮电大学,2009.
[7] 刘宗成,张忠林,田苗凤. 基于关联规则的网络行为分析[J]. 电子科技,2015,28(09):16-18,22.

(上接第 199 页)

[13] 邓子云,黄友森,杨晓峰,等. 基于 SOA-BPM 组合架构的第三方物流企业信息系统集成平台[J]. 计算机系统应用,2010,19(03):1-6.
[14] 尹裴,王洪伟,周曼. SOA 架构下面向业务敏捷性的信息系统柔性设计[J]. 情报杂志,2010,29(07):133-140.
[15] 倪枫. SOA 敏捷架构的 TOGAF 层次化迭代建模[J]. 上海理工大学学报,2018,40(04):364-370,390.
[16] 李润晔,倪枫,刘姜,等. 基于面向服务业务流程管理的系统架构建模[J]. 上海理工大学学报,2019,41(06):605-616.
[17] HACHICHA M, FAHAD M, MOALLA N, et al. Performance assessment architecture for collaborative business processes in BPM-SOA-based environment[J]. Data & Knowledge Engineering, 2016, 105(C):73-89.
[18] OMG. Business Process Model and Notation(BPMN) Version2.0.2[EB/OL]. [2017-09-09]. http://www.omg.org/spec/BPMN/2.0.2.
[19] 李昆颖,李效恋,张玲,等.基于 BPMN 的流程架构研究与实践[J]. 信息系统工程,2021(08):146-149.
[20] 凌晓东. SOA 综述[J]. 计算机应用与软件,2007,24(10):122-124,199.