



研究与开发

基于电信大数据的 5G 网络海量用户复访行为预测模型

孙玉娣

(江苏经贸职业技术学院数字商务学院, 江苏 南京 211168)

摘要: 5G 网络中的用户会产生大量的访问数据, 导致用户复访行为难以精准预测, 因此提出基于电信大数据的 5G 网络海量用户复访行为预测模型。从电信大数据中提取用户上网历史行为特征数据, 构建数据集。引入多阶加权马尔可夫链模型, 通过计算各阶自相关系数, 得到模型权重值, 计算模型的统计量。经过分析后得到各阶步长的马尔可夫链一步转移概率矩阵, 从而实现对 5G 网络海量用户复访行为的精准预测。实验结果表明, 该模型拥有最低的均值误差和标准差, 以及最高的精度、查全率、查准率、 $F1$ 指标, 可证明该方法在预测用户复访行为方面有着非常明显的优势。

关键词: 电信大数据; 用户复访行为预测; 多阶加权马尔可夫链模型; 一步转移概率矩阵; 自相关系数

中图分类号: TP357

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023026

A prediction model of massive 5G network users' revisit behavior based on telecom big data

SUN Yudi

School of Digital Commerce, Jiangsu Vocational Institute of Commerce, Nanjing 211168, China

Abstract: Users in 5G networks will generate a large amount of access data, which makes it difficult to accurately predict users' revisit behavior. Therefore, a prediction model of massive 5G network users' revisit behavior based on telecom big data was proposed. The user's historical online behavior characteristic data was extracted from the telecom big data to build a data set. Multi order weighted Markov chain model was introduced. The model weight value was obtained by calculating the autocorrelation coefficient of each order, and the statistics of the model were calculated. After analysis, the one-step transition probability matrix of Markov chain with each step size was obtained, so as to accurately predict the revisit behavior of massive users in 5G network. The experimental results show that the proposed model has the lowest mean error and standard deviation, as well as the highest accuracy, recall, precision and $F1$ indicators, which can prove that the proposed method has a very obvious advantage in predicting users' revisit behavior.

Key words: telecom big data, prediction of users' revisit behavior, multi order weighted Markov chain model, one step transition probability matrix, autocorrelation coefficient

收稿日期: 2022-12-28; 修回日期: 2023-02-07

基金项目: 2021 年江苏高校“青蓝工程”优秀教学团队项目; 江苏经贸职业技术学院“领军人才”资助项目

Foundation Items: “Qing Lan Project” in Jiangsu Universities in 2021, “Leading Talents” Program of Jiangsu Vocational Institute of Commerce



0 引言

随着 5G 电信网络迅速发展,人们可以通过各种各样的网站浏览新闻、下载数据以及购买商品,在方便生活的同时丰富了知识储备。这一系列操作必然会产生海量的网络数据,利用相关算法从这些数据中挖掘出有用的信息,并对用户未来可能访问的网站和购买的商品进行预测,已经成为一项十分热门的研究内容。针对可能复访或者复购的用户,根据其先前访问的历史和偏好进行针对性的推荐,可以在一定程度上提高用户的购买欲望。用户的浏览、操作、访问等历史行为数据都以日志文件的形式存储在数据库中,如何利用这些行为数据分析用户是否会复访,对于网络平台的可持续发展具有十分重要的意义。

文献[1]将深度神经网络算法与不用正则化方法联合起来,通过建立不同的分组,根据一定的数据特征对数据集进行复访行为的预测;文献[2]在用户行为序列的基础上实现用户点击预测。按照交互时间对用户历史行为进行排序,得到用户历史行为序列;将词嵌入模型引入深度因子分解机(deep factorization machine, DeepFM)模型,对用户历史行为序列进行自适应学习,得到用户的兴趣列表,捕捉用户的兴趣变化,从而实现预测。

上述两种方法已无法适应当前的 5G 大数据网络环境,因此,本文提出了一种基于电信大数据的 5G 网络海量用户复访行为预测模型。首先,从服务器节点中提取用户的浏览数据、行为数据、操作数据以及属性数据等各类信息构建 5G 电信网络数据集;然后,构建多阶加权马尔可夫链模

型,并对模型的转移矩阵和初始概率向量进行计算;最后,根据各阶步长的自相关系数计算权重值,分析权重值后得到各阶步长的马尔可夫链一步转移概率矩阵,实现对 5G 网络用户复访行为的精准预测。在实验中,将本文模型与其他方法进行预测性能对比,结果表明本文模型在多个方面均展现出了明显优势,预测均值误差、标准差始终低于其他两种方法,而预测精度则大大高于其他两种方法。

1 5G 网络用户复访行为预测

1.1 建立 5G 电信网络数据集

在进行用户复访行为预测之前,需要建立 5G 电信网络数据集^[3],为了确保用户行为数据的精准性和实时性,在 5G 电信网络中选取若干个服务器节点,将采集装置部署在这些节点上进行数据采集。采集内容包含用户浏览数据、用户属性数据、用户访问行为数据^[4]、用户访问深度数据等几大类数据,5G 电信网络采集数据解析见表 1。

5G 电信网络数据的采集频率^[5]设定为 0.2 次/s,根据采集信息种类的不同,将数据分别存储在 30 个数据库中,其中包含 280 多个字段以及若干个扩展字段。本文采集的数据来自真实网站的公开数据库,数据表示用户访问一次页面的所有浏览、操作行为,可以真实、有效地反映用户的行为特点。

本文数据采集的时间段为 2020 年 8 月—2021 年 6 月,其中包含 3 786 个用户 ID 和 38 468 次页面完整浏览记录。用 D 表示 5G 电信网络数据集, S 表示训练集, T 表示测试集。将 2020 年 8 月—2021 年 6 月的数据按照每两个月为一个单

表 1 5G 电信网络采集数据解析

数据类型	具体内容	数据来源
用户浏览数据	用户访问的页面总数、页面的用户访问量	用户行为分析日志
用户属性数据	用户账号、密码、单位、IP 地址、操作系统、终端	5G 网络服务器日志
用户访问行为数据	用户访问页面的频率、时间间隔、访问次数、停留时间以及页面被访问的平均次数	5G 网络服务器日志
用户访问深度数据	页面跳出率、用户访问页面的设备和方式	5G 网络服务器日志

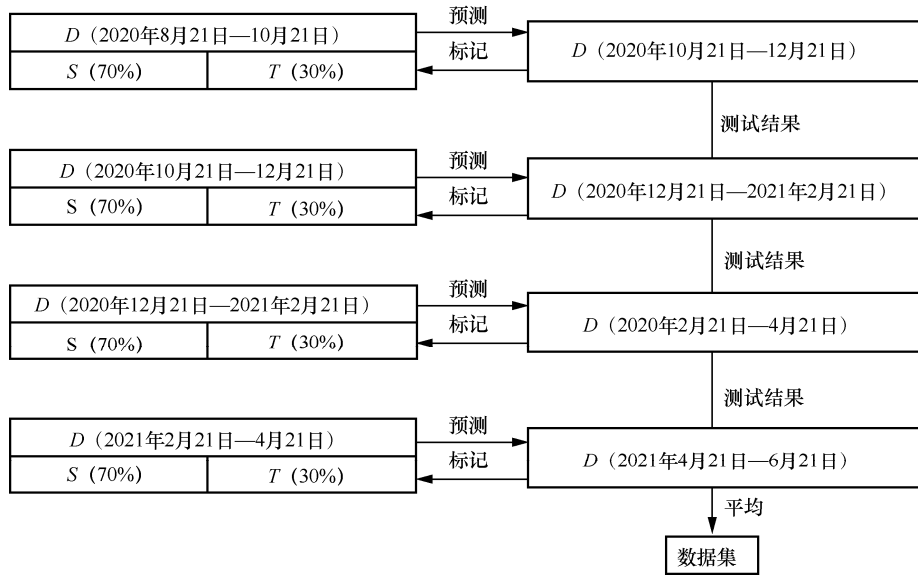


图1 5G 电信网络数据集构建过程

元的形式进行整合,共得到5个单元,作为数据集 D 。将上述所有用户数据的70%提取出来,作为训练集 S ,剩下的30%则作为测试集 T 。利用后两个月的用户复访行为^[6]对前两个月的数据进行标记,然后对总的结果进行平均,完成数据集的构建,5G电信网络数据集构建过程如图1所示。

1.2 用户复访行为预测模型

1.2.1 多阶加权马尔可夫链模型

由于电信大数据具有用户数量大、用户产生的数据量大、用户数据多样等诸多特点,在对其进行分析处理时常常出现效率低、难度大等问题。为此,引入马尔可夫链模型^[7-9],对5G电信网络用户进行复访行为预测。

马尔可夫链模型针对用户的上网行为做出了以下假设:用户上网浏览的过程是一个随机过程,即齐次的离散马尔可夫链,因此可以将用户上网行为构成的特征集合看作离散随机变量 $X^{[10]}$ 的值域,也就是说,用户上网过程构成了 X 的取值序列,且序列具有马尔可夫性。

根据上述假设构建马尔可夫链模型,以三元组 $M = \langle X, A, \alpha \rangle$ ^[11-13]的形式对马尔可夫链模型进行描述,其中, X 的值域为 $\{x_1, x_2, \dots, x_n\}$, x_n 对

应于用户的行为状态,也可被看作马尔可夫链模型当前的行为状态; A 代表转移概率矩阵,矩阵中的任意项 $p_{ij} = P(X_t = x_j | X_{t-1} = x_i)$ 都可以用来描述状态 x_i 向状态 x_j 转移的概率; α 表示马尔可夫链模型的初始状态分布^[14-15],分布中的项用 $p_i = P(X_{t=0} = x_i)$ 表示,即:

$$A = (p_{ij}) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots & p_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix} \quad (1)$$

$$\alpha = (p_i) = (p_1, p_2, \dots, p_n) \quad (2)$$

根据用户历史行为特征数据,整理出连续两个小时内出现用户上网行为的概率值 p_{ij} (前一段时间用户行为在 i 区范围内,之后则位于 j 区范围内),深入分析后可得到一步转移概率矩阵为

$$P_1 = (p_{ij}) \quad p_{ij} = \frac{N_{ij}}{\sum_{j=1}^n N_{ij}} \quad \text{且满足条件}$$

$\sum_{j=1}^n p_{ij} = 1 (i=1, 2, \dots, n)$,其中, N_{ij} 表示用户访问行为特征向量。那么,一步转移概率矩阵表达式为:



$$P_k = P_1^k \quad (3)$$

其中, k 表示矩阵维度。

将 $p(t) = (p_1(t), p_2(t), \dots, p_n(t))^T$ 看作 t 时刻下用户上网行为的绝对概率向量^[16-17], 其中, $p_n(t)$ 表示 t 时刻下用户上网行为位于 n 区的绝对概率。在式 (3) 的基础上, 进一步求取 $t+1$ 时刻下用户上网行为的绝对概率向量^[18-19]:

$$p(t+1) = p(t)P_k = p(t)P_1^k \quad (4)$$

那么, 在已知初始概率向量 $p(0) = (p_1(0), p_2(0), \dots, p_n(0))^T$ 的情况下, 就可以根据式 (3) 得到经过时间 t 后预测用户是否会复访的多阶加权马尔可夫链模型:

$$p(t) = p(0)P_t = p(0)P_1^t \quad (5)$$

综上所述, 只要已知马尔可夫链模型的初始概率向量, 就可以实现对任何时间下用户的复访概率以及复访网络区间的预测。

1.2.2 用户复访行为预测

(1) 从电信大数据中提取出用户上网历史行为特征数据, 并对各阶自相关系数 r_k ^[20-21] 和规范自相关系数 w_k 进行计算:

$$r_k = \frac{\sum_{i=1}^k (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^k (x_i - \bar{x})^2} \quad (6)$$

$$w_k = \frac{|r_k|}{\sum_{k=1}^m |r_k|} \quad (7)$$

其中, $\bar{x}$ 表示马尔可夫链模型的状态均值, m 表示根据实际预测需求计算得到的模型最大阶数。

对马尔可夫链模型各阶自相关系数做规范化处理^[22], 不同模型阶数下的 r_k 与 w_k 见表 2。

表 2 不同模型阶数下的 r_k 和 w_k

模型阶数	r_k	w_k
1	0.739 8	0.399 2
2	0.532 6	0.287 4
3	0.373 4	0.201 5
4	0.207 3	0.188 7

(2) 根据表 2 计算统计量:

$$\chi^2 = 2 \sum_{i=1}^4 \sum_{j=1}^4 p_{ij} \left| \lg \frac{p_{ij}}{i \cdot j} \right| \quad (8)$$

根据式 (8), 如果存在 $\chi^2 > \chi_\alpha^2((m-1)^2)$, 其中, χ_α 表示访问需求统计阈值, 则认为用户会出现复访行为, 否则认为用户不会出现复访行为。

2 实验测试

为了验证本文模型在实际应用中是否同样合理有效, 进行对比实验测试。实验所用数据从某大型网络的公开数据库中提取得到, 为了更好地进行实验, 预先对采集到的数据进行清洗处理, 剔除掉缺失率较大的缺失值, 并利用 scikit-learn 接口中的分类模型对数据集进行训练。

首先, 将本文模型与文献[1]和文献[2]提出的模型进行对比。分别应用 3 种模型对同一时间段内的用户上网行为进行分析, 并给出最终的复访行为预测结果。3 种模型的用户复访行为预测均值误差和标准差分别如图 2、图 3 所示。

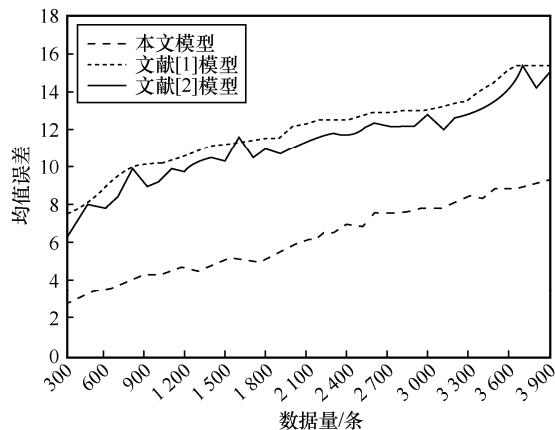


图 2 3 种模型的用户复访行为预测均值误差

通过观察图 2 和图 3 可以很清楚地看出, 随着数据量的不断增加, 本文模型的用户复访行为预测均值误差和标准差最小, 文献[2]模型的均值误差较文献[2]模型低一些, 而文献[1]模型的标准差较文献[2]模型低一些。

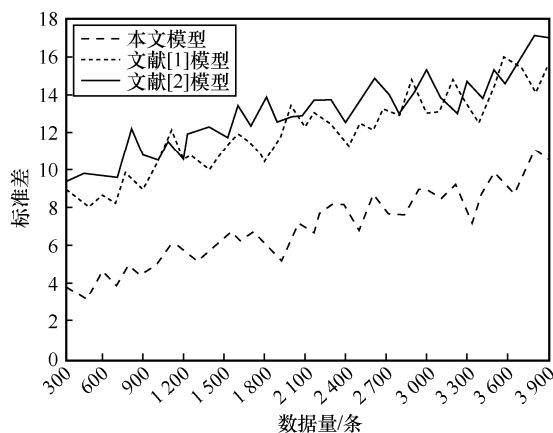


图3 3种模型的用户回访行为预测标准差

接下来通过查全率 R 、查准率 P 、 $F1$ 指标、精度 ACC 以及受试者操作特征 (receiver operator characteristic, ROC) 曲线下面积 (area under the curve, AUC) 5 个指标, 进一步验证 3 种模型的用户回访行为预测性能。用户回访行为预测从本质上来讲是一个二分类问题, 可以根据数据样本的真实类别和算法预测的类别将预测结果分为真阳性 (true positive, TP)、假阳性 (false positive, FP)、真阴性 (true negative, TN)、假阴性 (false negative, FN) 4 种。TP、FP、TN、FN 之和等于数据样本总数。当算法预测结果为 TP+FP、TP+FN 时, 表示正类; 当结果为 FN+TN、FP+TN 时, 表示负类。

查准率 P 和查全率 R 的表达式分别为:

$$P = \frac{TP}{TP+FP} \quad (9)$$

$$R = \frac{TP}{TP+FN} \quad (10)$$

P 与 R 是一对相互矛盾的指标, 如果要求确保算法的 R 值高就需要包含更多的正类, 这样一来就会造成算法的 P 值下降。

$F1$ 指标指的是 P 和 R 的调和平均值, 同时考虑了算法的 P 和 R 。

$$F1 = \frac{2 \times P \times R}{P + R} \quad (11)$$

ACC 是一个性能度量指标, 正确数据样本数量与数据样本总数的比值就是 ACC 。

$$ACC = \frac{1}{\beta} \sum_{i=1}^{\beta} I(x_i) \quad (12)$$

其中, β 表示实验数据总量, I 表示样本中的隐数据。

对训练集进行预测, 会得到一个预测概率, 将预测概率与概率阈值进行对比, 当预测概率大于阈值概率时, 数据样本为正类, 反之则被认定为负类。将训练集按照预测概率进行排序, 从而得到算法的最终预测性能。为了更加公平、准确地对比 3 种模型的预测性能, 引入 10 倍交叉验证法统计最终的实验结果, 3 种模型的用户回访行为预测结果见表 3。

表3 3种模型的用户回访行为预测结果

指标	本文模型	文献[1]模型	文献[2]模型
P	0.714 1±0.010 5	0.683 3±0.034 4	0.670 5±0.025 9
R	0.550 9±0.007 8	0.524 3±0.008 5	0.510 0±0.007 2
$F1$	0.621 9±0.001 0	0.594 3±0.018 0	0.547 1±0.017 3
ACC	0.701 8±0.021 0	0.683 8±0.006 1	0.663 1±0.001 3
AUC	0.686 9±0.001 3	0.655 9±0.008 3	0.642 9±0.004 5

通过观察表 3 可以看出, 3 种模型中, 本文模型的预测结果始终都是最优的, 由此可以说明本文模型在预测 5G 网络用户回访行为时的精准度最高。这是由于本文模型利用多阶加权马尔可夫链模型对电信大数据进行分阶分析和处理, 通过计算各阶步长的一步转移概率矩阵, 得到用户上网历史行为特征数据, 随着对特征数据分析的不断深入, 可得到用户回访行为预测结果。

3 结束语

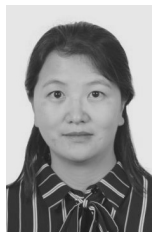
在 5G 电信网络环境下, 本文利用多阶加权马尔可夫链模型, 从大数据中提取用户上网历史行为特征数据, 通过对这些数据进行分析来确定用户的浏览习惯和偏好, 从而精准且高效地预测。将本文模型与其他模型进行对比实验, 实验结果表明, 本文模型有着最优秀的预测性能, 可实现对用户回访行为的精准预测。



参考文献:

- [1] 卢宇红, 宋佳丽, 王萌, 等. 基于深度神经网络融合稀疏分组 lasso 的预测模型研究[J]. 中国卫生统计, 2021, 38(6): 821-827.
LU Y H, SONG J L, WANG M, et al. The study on the prediction model based on deep neural network together with sparse group lasso[J]. Chinese Journal of Health Statistics, 2021, 38(6): 821-827.
- [2] 顾亦然, 王雨, 杨海根. 基于用户行为序列的短视频用户多行为点击预测模型[J]. 电子与信息学报, 2023: 10.11999/JEIT211458.
GU Y R, WANG Y, YANG H G. Multi-action click prediction model for short video users based on user's behavior sequence[J]. Journal of Electronics & Information Technology, 2023: 10.11999/JEIT211458.
- [3] CAO W C, WANG K, GAN H C, et al. User online purchase behavior prediction based on fusion model of CatBoost and Logistic[J]. Journal of Physics: Conference Series, 2021, 2003(1): 012011.
- [4] LI H R, LIN F Q, LU X, et al. Systematic analysis of fine-grained mobility prediction with on-device contextual data[J]. IEEE Transactions on Mobile Computing, 2022, 21(3): 1096-1109.
- [5] QIAO S B, PANG S C, WANG M, et al. Online video popularity regression prediction model with multichannel dynamic scheduling based on user behavior[J]. Chinese Journal of Electronics, 2021, 30(5): 876-884.
- [6] NIU B, SUI L, TANG J R, et al. Prediction of microblog users' forwarding behavior based on interactive and active information[C]//Proceedings of the 2020 International Conference on Aviation Safety and Information Technology. New York: ACM Press, 2020: 554-559.
- [7] XIAO Y P, LI J H, ZHU Y F, et al. User behavior prediction of social hotspots based on multimessage interaction and neural network[J]. IEEE Transactions on Computational Social Systems, 2020, 7(2): 536-545.
- [8] HU G Y, ZHOU Z J, HU C H, et al. Hidden behavior prediction of complex system based on time-delay belief rule base forecasting model[J]. Knowledge-Based Systems, 2020, 203: 106147.
- [9] SUDAN B, CANSIZ S, OGRETICI E, et al. Prediction of success and complex event processing in E-learning[C]//Proceedings of 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). Piscataway: IEEE Press, 2020: 1-6.
- [10] SOLTANI N Y. Online learning of sparse Gaussian conditional random fields with application to prediction of energy consumers behavior[C]//Proceedings of 2021 IEEE Statistical Signal Processing Workshop (SSP). Piscataway: IEEE Press, 2021: 486-490.
- [11] SUN L T, GAO S W, WANG L. An automatic test sequence generation method based on Markov chain model[C]//Proceedings of 2021 World Conference on Computing and Communication Technologies (WCCCT). Piscataway: IEEE Press, 2021: 91-96.
- [12] DENNIS L A, FU Y, SLAVKOVIK M. Markov chain model representation of information diffusion in social networks[J]. Journal of Logic and Computation, 2022, 32(6): 1195-1211.
- [13] PENG L, WEN L, QIANG L, et al. Research on complexity model of important product traceability efficiency based on Markov chain[J]. Procedia Computer Science, 2020, 166: 456-462.
- [14] HAN C, CHEN J, TAN M K, et al. A tensor-based Markov chain model for heterogeneous information network collective classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(9): 4063-4076.
- [15] CRUZ I R, LINDSTRÖM J, TROFFAES M C M, et al. Iterative importance sampling with Markov chain Monte Carlo sampling in robust Bayesian analysis[J]. Computational Statistics & Data Analysis, 2022, 176: 107558.
- [16] ALAMOUDI A, LIU M L, PAYANI A, et al. Predicting mobile users traffic and access-time behavior using recurrent neural networks[C]//Proceedings of 2021 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE Press, 2021: 1-6.
- [17] LIU K, TATINATI S, KHONG A W H. A weighted feature extraction technique based on temporal accumulation of learner behavior features for early prediction of dropouts[C]//Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE). Piscataway: IEEE Press, 2021: 295-302.
- [18] SETIA S, JYOTI V, DUHAN N. HPM: a hybrid model for user's behavior prediction based on N-gram parsing and access logs[J]. Scientific Programming, 2020: 1-18.
- [19] CHEN L Y, WANG L H, ZHOU Y X. Research on data mining combination model analysis and performance prediction based on students' behavior characteristics[J]. Mathematical Problems in Engineering, 2022: 1-10.
- [20] RASOULI A, ROHANI M, LUO J. Bifold and semantic reasoning for pedestrian behavior prediction[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 15580-15590.
- [21] ZHOU H, YU K M, CHEN Y C, et al. A hybrid feature selection method RFSTL for manufacturing quality prediction based on a high dimensional imbalanced dataset[J]. IEEE Access, 2021, 9: 29719-29735.
- [22] JIANG L, LIU H, JIANG H, et al. Heuristic and neural network based prediction of project-specific API member access[J]. IEEE Transactions on Software Engineering, 2022, 48(4): 1249-1267.

[作者简介]



孙玉娣 (1981—), 女, 江苏经贸职业技术学院数字商务学院副教授, 主要研究方向为本体、知识工程。