基于异构信息网络的学生成绩预测与预警模型研究*

徐小玉

(浙江万里学院 文献与信息中心,浙江 宁波 315000)

摘 要:推荐系统是数据挖掘中强有力的技术,异构信息网络是推荐系统起步晚却发展迅猛的主流推荐方法。提出了基于异构信息网络的学生成绩预测与预警模型,该方法通过元路径计算得到学生间相似度矩阵,利用相似度矩阵构造成绩变化趋势矩阵和幅度矩阵,投票得到学生成绩预警与预测结果;最后,在公开数据集上验证所提模型的有效性,结果表明,该模型能够对学生成绩进行预警,并能在一定阈值下预测学生成绩具体分值。

关键词:异构信息网络;元路径;成绩预警;成绩预测;数据挖掘

中图分类号:TP181 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2022.01.013

引用格式:徐小玉.基于异构信息网络的学生成绩预测与预警模型研究[J].信息技术与网络安全,2022,41(1):84-89.

Research on the prediction and early warning model of student achievement based on heterogeneous information network

Xu Xiaoyu

(Literature Information Center, Zhejiang Wanli University, Ningbo 315000, China)

Abstract: Recommendation system is a powerful technology in data mining. Heterogeneous information network is the mainstream recommendation method which started late but developed rapidly. In this paper, we proposed a student achievement prediction and early warning model based on heterogeneous information network. This method obtains the similarity matrix among students by Meta-Graph calculation, constructs the achievement change trend matrix and amplitude matrix by using the similarity matrix, and obtains the student achievement early warning and prediction by voting. Finally, the effectiveness of the proposed model is verified on the open data set, the experimental results show that the model can give early warning to the student achievement and can predict the specific score of students' performance under a certain threshold.

Key words: heterogeneous information network; Meta-Graph; academic warning; academic prediction; data mining

0 引言

教育数据挖掘旨在从海量的教育数据中发现隐藏在其中的内在联系与规律,为学生学习、教师教学以及教育管理提供一些帮助[1]。学生成绩预测与预警作为教育数据挖掘领域的重要研究分支之一,学生成绩预测与预警能帮助学生完善自我认知,提高自我学习能力,提升学生成绩及教师的教学成果,并且有助于教师对预警学生进行有效的干预和指导,具有重要的研究意义与应用价值。目前,对学生成绩进行预测分析及其成绩关键影响因素挖掘研究已引起国内外学者的关注,如张福生等的基于校园云的高校学生学业监测与预警系统研究[2],周庆的基于数据挖掘技术的高校学生学业预警分析研究[3],尹茂竹的基于大数据的高校学生学业成绩预警分析[4],李梦莹的基于双路注意力机制的学生成绩预测模型[5]。

在已有研究中,大多数学生成绩只是给予在成绩类别上的预测,如好、中、差等[5-8],少有能给出学生成绩分值的直接预测,而且鲜有学者利用异构信息网络对学生成绩预测与预警进行研究。本文提出了基于异构信息网络的学生成绩预测与预警模型,该方法通过元路径计算得到学生间相似度矩

*基金项目:宁波市教育科学规划课题(2021YGH038)

阵，利用相似度矩阵构造成绩变化趋势矩阵，投票得到学生成绩预警与预测结果；最后，在公开数据集上验证所提模型的有效性，结果表明，该模型能够对学生成绩进行预警，并能在一定阈值下预测学生成绩具体分值。

1 异构信息网络

异构信息网络(Heterogeneous Information Network，HIN)由 Sun[9]等人在 2009 年提出，异构信息网络推荐是大数据时代数据挖掘的主流方法。异构信息网络可以承载网络中的多种节点类型和节点之间的多种关联类型，能更加精准地定义出信息网络中的不同语意从而挖掘出更深层次的信息。异构信息网络在当前大数据时代有着广泛的应用[10-11]，例如交友网站、医疗信息网站、电子商务网站以及新闻网站等，其将复杂场景里的对象和对象之间的关联关系装载进一个结构化的网络并且通过该网络进行分析。

异构信息网络研究近十年的发展历程可以分为三个阶段：第一个阶段是起步阶段，该阶段研究主要集中在聚类、分析等方面[12-13]。第二个阶段是基于“元路径”进行相似度计算的异构信息网络阶段，在该阶段，Sun 等人提出了 PathSim 算法[14]，该算法用于对称元路径的相似性计算，例如在文献信息网络中查找两个作家在同一家出版社出版的图书可以用 PathSim 算法；Shi 等人提出了 HeteSim 算法[15]，该算法用于不同类型对象间相似关系度量。第三个阶段的研究方向主要是利用加权元路径进行相似度计算的异构信息网络[16-17]。

异构信息网络区别于现实世界网络，它是一种逻辑网络，该网络显示不同类型对象和不同类型关系的链接，如图书文献网络、购物网络、社交短视频网络等。由于其中的链接至关重要而且很难掌握被搜索对象的特征规律，因此这些网络有效搜索功能非常重要，其中比较特别的是为同类型对象找到类似的搜索内容，例如在社会媒介网络中，查找与某篇文章类似的文章。

本节主要对异构信息网络相关概念进行介绍，重点介绍对同类型进行度量的 PathSim 算法，并定义在教育数据方面的异构信息网络，给出教育数据异构信息网络的元路径定义。

定义 1[11] 信息网络(information network)是一个带有对象类型映射函数 $\varphi:V\rightarrow A$ 和链接类型映射函数 $\phi:E\rightarrow R$ 的有向图 $G(V,E)$，其中，任意 $v\in V$ 是一个不同的类型，记为 $\varphi(v)\in A$，每个链接 $e\in E$ 是一个特定关系类型 $\phi(e)\in R$，当 $|A|>1$ 或 $|R|>1$ 时，称该网络为异构信息网络，否则称为同构信息网络。

定义 2[11] 网络模式(network schema)是带有对象类型映射 $\varphi:V\rightarrow A$、链接映射 $\phi:E\rightarrow R$ 的异构信息网络 $G(V,E)$(G 是定义在对象类型集合 A 和关系类型集合 R 上的有向图)的元模板，记为 $\mathrm{TG}(A,R)$。

异构信息网络的网络模式刻画了对象类别及其交互关系，并不简单关注网络中对象类型属性，其阐明网络中对象集合及对象间关系集合的类型限制，类似于数据库系统中的 ER。异构信息网络更能反映人类社会的交互活动，故而利用该网络研究学生间的相似性更为贴切。研究基于异构信息网络的推荐对于满足用户的个性化需求和多样化需求是十分必要的。

文献信息网络是常见的异构信息网络，其结构如图 1 所示。

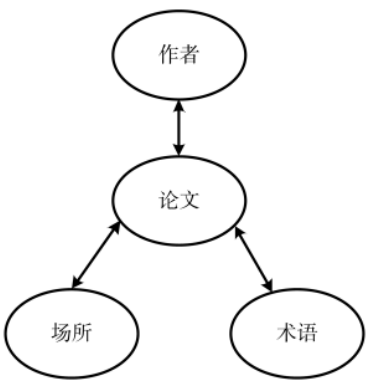

图 1 文献信息网络

在 HIN 中的两个对象能通过不同的属性类型相互连接，这些不同的属性路径具有不同含义。因此，两个对象的相似度依赖于异构信息网络的搜索路径。

定义 3[11] 元路径 P 是定义在网络模式 $\mathrm{TG}(A,R)$ 图上的一条路径，符号表示为：$A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\xrightarrow{R_l}A_l$(简记为 $P=A_1A_2\cdots A_l$)。

元路径的提出为 HIN 中对象间相似性度量及网络推荐提供了基础，本文所研究学生成绩预测与预警模型是基于对称元路径的相似性度量 PathSim 算法。

定义 4[14] PathSim：基于元路径的相似性度量，

给定一条对称的元路径 P，两个同类型 x 和 y 的 PathSim 是：

$$S\{x,y\}=\frac{2\times|\{P_{x\to y}:P_{x\to y}\in P\}|}{|\{P_{x\to x}:P_{x\to x}\in P\}|+|\{P_{y\to y}:P_{y\to y}\in P\}|}$$

其中，$P_{x\to y}$，$P_{x\to x}$，$P_{y\to y}$ 分别是 x 和 y 之间，x 和 x 之间，y 与 y 之间的路径实例。

2 学生成绩预测与预警模型

在信息时代下，各个学校基本都已建立了自己的服务系统以适应学生的需求，也都积累了海量的数据，如学生的生源地数据、科研数据、成绩数据、上网数据、晨跑数据等，这些数据刻画着学生生活、学习各个方面。

异构网络中多种节点类型和节点之间的多种关联类型，与当前社会和人类的交互活动更为相似，更能反映现实世界。利用异构信息网络刻画各种影响学生学习的因素关系，推荐出有相同影响因素的学生，通过这些学生成绩变化情况，给出与其相似的学生成绩预警或具体成绩值的预测。

2.1 问题定义

给定由教育数据构成的异构信息网络 DS=(V，E)，其中 V 表示网络中的节点，节点类型包括基本信息、学生、家庭因素、分数、在校表现、学期；E 表示网络中的边。学生教育数据异构信息网络模式如图 2 所示。

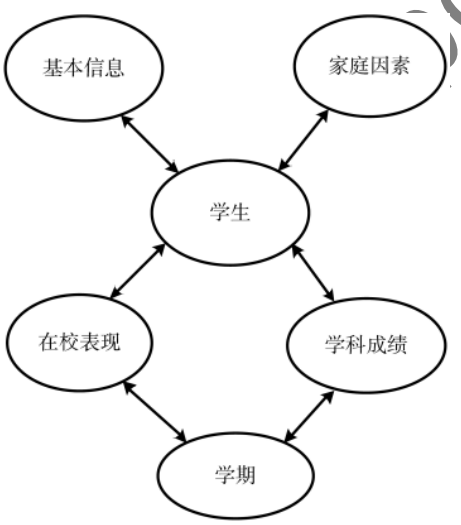

图 2 教育数据信息网络模式

为了更好地理解算法 PathSim 的原理以及初步解释所定义的教育信息网络，本文结合具体的例子进行说明。

例 1 在教育数据上一个简单含有三个实体项的学生成绩异构信息网络，实体对象类型为学生、学科成绩、学期，关系类型为学生和学科成绩之间是考取和被考取的关系，学科成绩和学期是属于和被属于的关系，其结构图如图 3 所示。以对称元路径 SGTGS 为例，它表示的语义是两个学生(S)在同一学期(T)的学科成绩(G)。表 1 是一个网络中学生和学期间复合计算后的邻接矩阵 W_{ST}，表示学生在每个学期所获得的成绩分值。

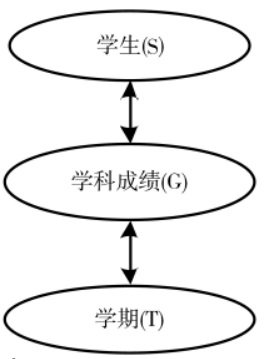

图 3 学生简单成绩信息网络模式

表 1 邻接矩阵

	G1	G2	G3
S1	13	10	11
S2	9	9	9
S3	13	11	11
S4	6	6	4

此例中关系矩阵 $M=W_{ST}W_{TS}$，计算结果如表 2 所示。

表 2 关系矩阵

	S1	S2	S3	S4
S1	390	306	400	182
S2	306	243	315	144
S3	400	315	411	188
S4	182	144	188	88

通过 PathSim 算法的相似度量公式计算可得：

$$S(S1,S1)=\frac{2\times390}{390+390}=1$$

$$S(S1,S2)=\frac{2\times306}{390+243}=0.966$$

$$S(S1,S3)=\frac{2\times400}{390+411}=0.998$$

$$S(S1,S4)=\frac{2\times182}{390+88}=0.761$$

故学生 S3 与学生 S1 在三个学期里成绩相似

性最高，从邻接矩阵 W_{ST} 中也能观察出学生 S3 与学生 S1 成绩分值最为接近。

2.2 模型的构造

本文基于学生教育数据提出一种基于异构信息网络的学生成绩预测与预警模型(HIN_SFY)，模型框架如图 4 所示。

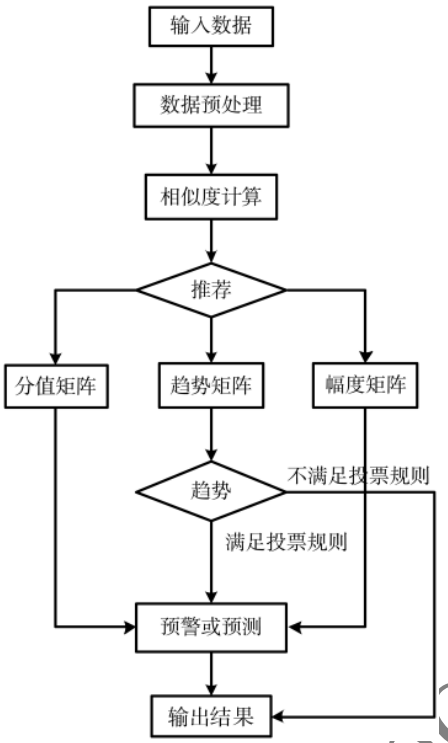

图 4 HIN_SFY 框架图

该模型共包含 3 层：

(1)输入编码层

首先对各属性值及历史成绩进行预处理，包括数值转换、归一化、分组等，在此基础上，生成各个属性的特征表示和历史成绩的特征表示。

(2)PathSim 相似层

利用 PathSim 算法计算得到学生在同一学期学习影响因素相似度矩阵 Ms：

$$Ms=\begin{bmatrix} s(a_1,a_1) & s(a_1,a_2) & \cdots & s(a_1,a_n) \\ \vdots & \vdots & \vdots & \vdots \\ s(a_n,a_1) & s(a_n,a_2) & \cdots & s(a_n,a_n) \end{bmatrix}$$

其中 $s(a_i,a_j)$ 为学生 a_i,a_j 的相似度。

在一定阈值下，选取相似度较高的学生成绩值构成学生成绩预测或预警分值矩阵 Fs、趋势矩阵 Qs 及幅度矩阵 Bs。

(3)预警和预测层

通过趋势矩阵 Qs 得到被预测学生成绩趋势，如增、平、减、平升、平降等，以及变化幅度，然后给出被预测学生在一定阈值下的成绩，若达到预警值则给出预警，提醒其端正态度、努力学习。

投票规则如下：

(1)若趋势矩阵中存在与被预测学生成绩变化趋势相同的学生，则根据他们在趋势矩阵的变化趋势给出被预测学生成绩变化趋势，并通过幅度矩阵中变化幅度给出幅度变化情况，最后给出学生成绩预警或成绩预测在一定阈值下具体分数值。

(2)若趋势矩阵中没有与被预测学生成绩变化趋势相同的学生，则将趋势矩阵中变化频率最高的趋势作为被预测学生成绩变化趋势，并给出相应预警或成绩分值预测。

3 实验

本文实验数据来自 kaggele 网站下 Studentperformance 中数据成绩数据集，内含有效数据 357 条，含有 30 个特征属性，3 个期末成绩。基于第 2 节所提的教育数据异构信息网络，数据集划分为基本信息、学生、家庭因素、分数、在校表现、学期这几个节点。随机抽取 80%的数据作为训练集，20%的数据留作测试集。

实验采用 SQLServer 数据库与 Python 语言结合，首先对数据进行预处理，利用异构信息网络的 PathSim 算法得到相似度矩阵，并得到趋势矩阵和幅度矩阵，最后在趋势矩阵和幅度矩阵上投票得到学生成绩预测与预警情况。实验中训练数据共 286 条，测试数据 71 条，设定 9 分为预警分数线，预测分数误差阈值为 1，实验结果显示预测正确 55 条，正确率为 77.4%。

鉴于数据属性较多，现取测试集中学生 A 和学生 B 的成绩预测过程(例 2、例 3)，学生 C 的成绩预警过程(例 4)来展示实验过程，为了方便预测和预警理解，表 3 给出三位学生原始成绩。

例 2 对于学生 A，通过 PathSim 算法计算得到与

表 3 学生原始成绩

	G1	G2	G3(预测)
学生 A	15	13	13
学生 B	14	15	16
学生 C	9	8	8

其学习影响因素相似度高的前 10 名学生，并用他们的三个学期期末成绩构成学生成绩预测或预警分值矩阵 Fs_A 和趋势矩阵 Qs_A（当然与学生 A 在校学习影响因素相似度最高的是其本身，这里将学生 A 被预测成绩和变化幅度用 x 和 y 代替，学生 B 和学生 C 预测或预警过程也类似设置）：

$$Fs_A=\begin{bmatrix}15&13&x\\14&14&14\\9&9&9\\15&12&12\\14&15&16\\10&10&10\\11&10&10\\12&10&10\\12&15&15\\13&14&13\end{bmatrix},Qs_A=\begin{bmatrix}-1&y\\0&0\\0&0\\-1&0\\1&1\\0&0\\-1&0\\-1&0\\1&0\\1&-1\end{bmatrix}$$

注：Qs_A 矩阵中负数代表下降，正数代表上升，0 代表持平。

先通过趋势矩阵 Qs_A 第一列可知与学生 A 在前两学期变化趋势相同的学生分别为学生 A4、学生 A7 和学生 A8，并且这三个学生第三学期的成绩变化值为零，则可知被预测学生第三学期的成绩变化趋势为平，即成绩分数值为 13，该结果与学生 A 原成绩相同。与 A 在前两学期变化趋势相同的学生 A4、学生 A7 和学生 A8 的成绩分数具体值如图5所示。

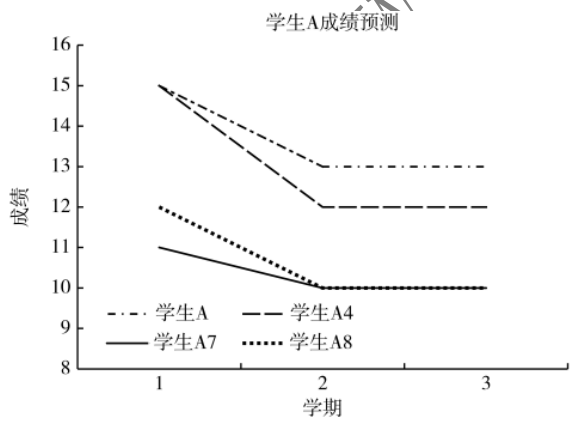

图 5 学生 A4、A7、A8 各学期成绩

例 3 对于学生 B，通过 PathSim 算法得到与其在校学习影响因素相似度高的前 10 名学生构成分值矩阵 Fs_B、趋势矩阵 Qs_B 及幅度矩阵 Bs_B：

$$Fs_B=\begin{bmatrix}14&15&x\\6&6&4\\17&17&17\\7&7&8\\9&9&9\\12&12&13\\13&12&12\\10&9&9\\12&12&12\\11&9&10\end{bmatrix},Qs_B=\begin{bmatrix}1&y\\0&-2\\0&0\\0&1\\0&0\\0&1\\-1&0\\-1&0\\0&0\\-2&1\end{bmatrix},Bs_B=\begin{bmatrix}x\\0\\1\\0\\1\\0\\0\\0\\1\end{bmatrix}$$

从趋势矩阵 Qs_B 第一列可知矩阵中没有学生成绩变化与学生 B 相同，则统计趋势矩阵第二列得到学生 B 第三学期的成绩变化趋势大多数为平或升，通过 Bs_B 矩阵可知升的幅度为 1，即给出的预测值为 15 或 16，与学生 B 实际成绩做对比可知，该值在误差阈值允许范围内，预测正确。

通过例 2 和例 3 可知，本模型可以预测出学生成绩具体分值，而不是把成绩划分类别，仅预测出学生成绩类别。

例 4 已知成绩预警线为 9 分，现有学生 C 前两学期成绩分别是 9 分和 8 分，对于学生 C 通过 PathSim 算法得到与其在学习影响因素相似度高的前 10 名学生，分别构成分值矩阵 Fs_C 和幅度矩阵 Qs_C：

$$Fs_C=\begin{bmatrix}9&8&x\\6&5&5\\8&9&10\\8&12&12\\6&8&8\\13&13&12\\8&8&10\\15&14&14\\8&7&6\\10&9&11\end{bmatrix},Qs_C=\begin{bmatrix}-1&y\\-1&0\\1&1\\4&0\\2&0\\0&-1\\0&2\\-1&0\\-1&-1\\1&2\end{bmatrix}$$

从趋势矩阵 Qs_C 第一列可知矩阵中学生 C 与学生C2、C8、C9 成绩变化趋势相同，且 C2、C8 第三学期成绩变化趋势为平，C9 第三学期成绩变化趋势为降，且三位同学有两位成绩低于 9 分，结合学生 C 本身成绩，需要给出该生成绩挂科预警。

将本文所提出的基于异构信息网络的学生成绩预测与预警方法(HIN_SFY)同基于双路注意力机制的学生成绩预测方法(TWA)[5]和传统的分类预测方法如支持向量机(SVM)、逻辑回归(LR)、高斯朴素

贝叶斯(NB)、决策树(DT),分别在 Studentperformance 数据集中进行对比实验,验证本文提出方法的有效性。实验结果如表 4 所示。

表 4 算法对比结果

	成绩类别预测	成绩分值预测
TWA	是	否
SVM	是	否
LR	是	否
NB	是	否
DT	是	否
HIN_SFY	是	是

4 结论

学生成绩预测与预警是目前教育大数据研究的热门领域,也是数字校园建设的重要组成部分。通过对学生成绩做预测和预警,一方面可以增强学生自我学习意识,另一方面也有利于教师管理和帮助学生;利用异构信息网络刻画各种影响学生学习的因素关系,推荐出有相同影响因素的学生,通过这些学生成绩变化情况,既能给出学生成绩预警并能在一定阈值下预测成绩具体分值。未来研究工作中,可以考虑采用异构信息网络的其他算法对学生成绩进行预测与预警,在实验中分析对比,找到预测正确率更高的算法。

参考文献

[1] 王盛.教育数据挖掘促进高校学生个性化学习途径分析[J].考试周刊,2014(34):176.

[2] 张福生,吕开东,韩伊佳.基于校园云的高校学生学业监测与预警系统研究[J].中国教育信息化,2015(9):87-89.

[3] 周庆,肖逸枫.基于数据挖掘技术的高校学生学业预警分析[J].中国教育技术装备,2018(6):36-39.

[4] 尹茂竹.基于大数据的高校学生学业成绩预警分析[D].天津:天津商业大学,2018.

[5] 李梦莹,王晓东,阮书岚,等.基于双路注意力机制的学生成绩预测模型[J].计算机研究与发展,2020,57(8):1729-1740.

[6] 吴晓倩,权丽丽,陈诚,等.基于大数据决策树算法的学生成绩分析与预测模型仿真[J].电子设计工程,2020,28(24):138-141,146.

[7] 刘爱萍.基于数据挖掘技术的高校学生成绩预测模型构建[J].长春工程学院学报(自然科学版),2020,21(2):98-101.

[8] 黄建明.贝叶斯网络在学生成绩预测中的应用[J].计算机科学,2012,39(S3):280-282.

[9] SUN Y,HAN J,ZHAO P,et al.RankClus:integrating clustering with ranking for heterogeneous information network analysis[C]//Proceedings of the 12th International Conference on Extending Database Technology,2009:565-576.

[10] 林怿星,唐华.基于异构信息网络的混合推荐模型[J].计算机应用,2021,41(5):1348-1355.

[11] 刘云枫,孙平,葛志远.异构信息网络推荐研究进展[J].情报学,2020,38(6):151-157.

[12] SUN Y,YU Y,HAN J.Ranking-based clustering of heteroge neous information networks with star network schema[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Paris,France.DBLP,2009:797-806.

[13] SUN Y,AGGARWAL C C,HAN J. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes[J].Proceedings of the VLDB Endowment,2012,5(5):394-405.

[14] SUN Y,HAN J,YAN X,et al.PathSim:meta path-based top-K similarity search in heterogeneous information networks[J].Proceedings of the VLDB Endowment,2011,4(11):992-1003.

[15] SHI C,KONG X,HUANG Y,et al.HeteSim:a general frame work for relevance measure in heterogeneous networks[J].IEEE Transactions on Knowledge & Data Engineering,2014,26(10):2479-2492.

[16] SUO X,WEI F,YU K.Entity recommendation via integrating multiple types of implicit feedback in heterogeneous information network[C]//IEEE International Conference on Data Mining Workshops.IEEE Computer Society,2017:781-786.

[17] VAHEDIAN F,BURKE R,MOBASHER B.Weighted random walks for meta- path expansion in heterogeneous networks[C]//RecSys 2016 Poster Proceedings,2016:15-19.

(收稿日期:2021-10-15)

作者简介:

徐小玉(1990-),女,硕士研究生,工程师,主要研究方向:人工智能、推荐系统。

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所