文章编号:1000-2367(2023)04-0043-09 DOI:10.16366/j.cnki.1000-2367.2023.04.006

基于多维用户画像和 DeepFM 的“环评云助手”资源推荐研究

李天玉[1],车蕾[1],丁峰[2],谭悦[1]

(1.北京信息科技大学 信息管理学院,北京 100192;2.北京尚云环境有限公司,北京 102208)

摘 要:“环评云助手”是一款服务于环境影响评价行业用户的 APP,针对 APP 中信息量激增、行业资源文本特征利用不充分和行业用户即时资源推荐精准较低等问题,提出一种结合行业文本资源和用户行为特征的多维用户画像模型并应用于深度因子分解机(Deep Factorization Machines,DeepFM),实现资源点击率(Click-Through-Rate,CTR)的精准预测.模型首先对行业资源文本进行语义抽取,再对行业用户行为进行自定义评分,从而构建多维用户画像模型;最后将多维用户画像应用于 DeepFM 模型,进行 CTR 预测任务,实现具有行业特征的个性化推荐.实验数据来自“环评云助手”APP,实验结果表明该模型有效提高了 CTR 预测任务的 AUC 值,降低了 LogLoss 值,具有一定的应用价值.

关键词:环境影响评价;用户画像;标签生成;推荐算法;深度因子分解机;CTR 预测任务

中图分类号:TP391.1 **文献标志码**:A

环境影响评价(以下简称“环评”)可定义为:对规划和建设项目实施后可能造成的环境影响进行分析、预测和评估,提出预防或者减轻不良环境影响的对策和措施.环评行业用户画像是画像技术在环境评估领域的具体应用,它在用户画像的基本理念上添加了新的约束条件和应用场景.在大数据时代背景下,用户信息分散,面对如此丰富的海量数据,将用户信息抽象成标签,加以组合利用,挖掘出隐藏在大数据中的信息可以为用户提供更加精准的、有效的个性化服务.近年来,用户画像在推荐算法领域取得了系统性的突破,但基于环评行业特征来解决该行业用户间资源推荐的研究还有待深入,存在的一些问题还有待去解决.

“环评云助手”是一款服务于环境影响评价行业用户的 APP,其主要功能包括标准政策查询浏览、分类管理名录查询等,包含国家和地方发布的法律法规、政策文件、标准规范等 22 000 余条,100 000 余名环评从业者或行业业余人员注册使用,月活跃度高达 40 000 余人.

本文充分利用“环评云助手”的行业文本资源和行业用户行为特征,构建体现行业特征的用户画像模型;同时结合深度因子分解机模型,以提高“环评云助手”资源推荐性能,满足平台用户精准获取有用资源的需求.模型在泛化能力和适用能力等方面都有相应提升.本文主要贡献度如下:

(1)更有效地利用平台行业文本资源和用户行为特征.模型同时考虑行业文本资源中长短文本对用户画像、标签的贡献性,并通过自定义规则对用户行为进行评分,多维挖掘行业特征.

(2)将用户画像与 DeepFM 模型结合,更准确地预测资源点击率(CTR),以提高算法的推荐效率和综合评价指标.

(3)模型在“环评云助手”数据集上进行实验并取得了很好的效果.开展与其他模型的对比实验,实验结果表明,模型在各评价指标方面均优于其他模型.

本文接下来首先阐述相关研究工作,第 2 节深入探讨行业用户画像模型的构建,第 3 节探讨将用户画像应用于 DeepFM 模型,第 4 节展示并分析实验工作及结果,最后对全文进行总结并对该研究方向进行展望.

收稿日期:2022-06-07;**修回日期**:2022-06-29.

基金项目:国家自然科学基金(51975058);教育部人文社科规划基金(20YJAZH129);2022 年北京信息科技大学优质课程专项;北京信息科技大学 2023 课程思政立项项目(2023JGSZ20).

作者简介:李天玉(1996—),女,北京市人,北京信息科技大学硕士研究生,研究方向为用户画像、推荐系统.

通信作者:车蕾(1979—),女,河南洛阳人,北京信息科技大学副教授,博士,研究方向为深度学习、自然语言处理,E-mail:che.lei@163.com.

1 相关研究工作

用户画像是从海量信息中抽取出用户信息的集合,用于描述用户需求、偏好与兴趣的模型[1].最早提出用户画像概念的是交互设计之父 A. Cooper,他将用户画像定义为"基于用户真实数据的虚拟代表". QUINTANA 等[2]也将用户画像描述为"一个从海量数据中获取并由用户信息构成的标签集合",通过这些标签信息,可以反映用户的需求、个性化偏好等.用户画像方法虽然起源于公安情报,在电子商务领域得到壮大发展,但如今在图书情报[3]、科技情报[4]、社交论坛等领域都发挥着重要作用.当前,面向基于实证研究平台的环评行业画像研究仍是一个较为全新的领域,通过梳理画像技术在用户画像领域的发展,可以为环评行业画像的研究和应用提供借鉴.

20 世纪 90 年代,协同过滤技术的首次提出[5],标志着推荐系统成为一门独立的学科而受到广泛关注.如今,许多学者都在传统推荐模型的基础上结合用户标签特性和用户画像技术提出了新的个性化推荐方法.张亮[6]融合用户、标签、资源,利用 LDA 构建主题模型,通过融合对象间关系与资源内容特征进行标签推荐.熊回香等[7-9]在此研究基础上,不仅提出了从资源-标签-用户 3 个维度分别建立推荐组件,还构建了基于社会化标签的单用户和群用户兴趣模型,通过协同过滤算法的思想,架构了个性化信息服务流程.李兴华等[10]提出了基于兴趣-标签的 ITRA 推荐算法,将用户候选兴趣集、推荐兴趣-标签集、项目推荐集作为最终的推荐结果.

CTR 预估用来估计用户点击推荐资源的概率,在推荐系统中极为重要.对于一个基于 CTR 预估的推荐系统,重要的是学习到用户行为潜在的特征组合.在不同的推荐场景中,低阶组合特征或高阶组合特征都可能会对最终的 CTR 预测结果产生影响.因子分解机(Factorization Machines,FM)是经典的 CTR 预估模型,通过对每一维特征的隐变量内积来提取特征组合,从而进行点击率预测,但是 FM 因为计算复杂度等原因只用到了二阶特征组合,不能获得高阶特征交互.为了解决上述问题,JUAN 等[11]在 FM 的基础上引入 field 的概念,提出了领域知识因子分解机模型(Field-aware Factorization Machine,FFM),将每个 field 的 embedding 值传入 MLP,从而获取了高阶特征交互.2017 年,GUO 等[12]为了减少 Wide&Deep 模型中的特征工程,提出了 DeepFM,将 embedding 后的特征表示同时传入浅层网络和深层网络,通过端到端的方式同时获得了浅层特征交互表示与深层特征交互表示.

由于上述文献方法缺少行业特征的渗透,若直接应用在"环评云助手"APP 中,将很难精准构建用户画像并准确预测 CTR 点击率,以满足环评行业用户的资源推荐需求.因此,本文结合行业特征,提出了一种融合文本资源特征和用户行为特征的画像模型并结合 DeepFM 模型实现用户个性化推荐.

2 "环评云助手"多维用户画像构建

基于 APP 数据集特征,先后提取"环评云助手"文本资源特征标签和用户行为特征进行自定义评分,并通过这两个维度构建环评行业用户画像要素关联路径,进而构建"环评云助手"多维用户画像模型.

2.1 基于文本资源特征的标签集构建

本文基于环评行业文本资源特征,从标题短文本和摘要长文本两方面进行考虑,多维度构建用户画像.从逻辑结构来看,文本标题属于短文本,具有揭示环评资源内容主旨的作用;文本摘要属于长文本,阐明了该资源的适用范围及主要内容.这两种文本在挖掘行业特征方面都起到重要作用,不仅能从行业文本资源特征中发掘用户兴趣,也充分考虑了文本逻辑结构对画像模型构建的影响.

2.1.1 基于标题短文本的标签构建

基于标题短文本的画像标签融合了行业词、关键词和主题词三方面.将行业词记作 L_{industry},关键词记作 L_{key},主题词记作 L_{topic},共计 m 个用户,则第 i 个用户 u_i 基于标题短文本的画像标签为:

$$L_i=[L_{\text{industry}_i},L_{\text{key}_i},L_{\text{topic}_i}].$$

(1)基于标题短文本的行业词.《建设项目环境影响评价分类管理名录》(以下简称《分类管理名录》)是环境影响评价领域重要的参考指标.该名录划分了 55 个一级分类,如农业、林业、畜牧业、渔业等;一级分类中

又下分了 173 个小类,例如畜牧业类中包括了牲畜饲养、家禽饲养和其他畜牧业.本文统计了资源的分类名录信息作为该资源的行业词,一定程度上体现了用户较为关注和感兴趣的行业领域.

(2)基于标题短文本的关键词.使用词频-逆文档频率(Term Frequency-Inverse Document Frequency,TF-IDF)算法进行词频统计,计算每个候选关键词的综合权重,从而依据该权重值对候选关键词进行排序,得到高权重的关键词[13].对资源标题文本使用此方法不仅可以生成作为标签的词汇,还反映该用户在环评行业中最关注的领域关键词.例如,某用户的关键词中,出现"水质"的比例远远高于其他词汇,则考虑该用户在环评行业中对水质领域的关注程度较高、从事水质方面工作的可能性较大.

(3)基于标题短文本的主题词.隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)主题模型将文档的主题以概率分布的形式给出,从而通过分析文档、抽取主题分布后进行主题聚类.环评行业资源的标题文本具有一定的专业性和结构性,可使用 LDA 主题模型对资源标题文本进行主题聚类,得到每个主题下的行业主题词.例如,一些标题文本中会出现"水质、光谱法、污染物、排放……"等围绕环评方面的专业词,且该领域的专业划分明确,由此可以通过 LDA 主题模型生成围绕环评行业主题展开的主题词.

2.1.2　基于摘要长文本的特征提取

基于摘要长文本的特征提取,其目的要抽取资源摘要中的文本特征,该方法使用 TextRank 文本摘要抽取算法,衡量每个句子与其他句子之间的联系,求出该句子的候选权重,从而抽取主要内容作为候选句[14].将用户记作 u_i,候选句权重记作 w_i,候选句记作 c_i,则摘要生成结果根据候选权重 w_i 排序,结果记为 $L_{\text{abstract}_i}=[c_{i,1},c_{i,2},c_{i,3}]$. 其主要 5 个步骤如下所示:

(1)对文本 T 进行句子分割,即 $T=[S_1,S_2,\cdots,S_n]$;

(2)对每个句子 $S_i\in T$,进行分词,停用词、无意义的词过滤等操作,即 $S_i=[t_{i,1},t_{i,2},\cdots,t_{i,n}]$;

(3)识别文本单元之间的关系,分别添加到图模型中形成节点和边;

(4)对各节点的权重进行迭代计算,直到计算结果收敛,其公式如下所示:

$$WS(V_i)=(1-d)+d\sum_{V_j\in \text{In}(V_i)}\frac{w_{ji}}{\sum_{V_K\in \text{Out}(V_j)}w_{jk}}\times WS(V_j),\tag{1}$$

其中,$\text{In}(V_i)$ 表示指向节点 V_i 的节点集,$\text{Out}(V_j)$ 表示指向节点 V_j 的节点集,w_{ji} 表示节点 V_j 指向节点 V_i 的边权重,d 表示阻尼系数,通常取 0.85;

(5)对候选句权重倒序排序,将权重排序中前 3 个句子作为目标文本的摘要句,若目标文本中的候选句数量小于 3,则选取当前全部候选句作为摘要结果 $L_{\text{abstract}_i}=[c_{i,1},c_{i,2},c_{i,3}]$.

2.2　基于用户行为的评分矩阵构建

用户行为评分,可以将用户与资源的交互行为数值化,体现了用户对资源的兴趣程度.所以通过统计用户与资源之间的交互行为,分析其行为轨迹,建立行为轨迹与资源评价的关系,把用户对资源的交互行为转换成对应的兴趣评分,不仅挖掘了用户感兴趣的资源,也在一定程度上改善了算法的矩阵稀疏问题[15].

本文从用户对环评行业文本资源的浏览、收藏、分享和评价行为入手,分别统计用户对资源的浏览次数、评论次数、分享次数与收藏情况.本文采用自定义评分规则,参考付芬等[16]和顾寰等[17]对用户行为评分的定义规则,定义评分取值范围为 $R_{jk}\in[0,5]$.具体分值定义规则依据"环评云助手"用户等级加分规则和 APP 虚拟货币"云贝"累计加分规则,各项评分由这两方面加权平均得到.具体评分规则如表 1 所示.

表 1　用户行为评分标准表

Tab. 1　User behavior scoring criteria table

用户交互行为	行为变量	用户等级加分	"云贝"加分	评分
浏览	r_{browser}	1	1	1
收藏	r_{collect}	2	0	1
分享	r_{share}	2	0	1
评论	r_{comment}	3	1	2

(1)定义 R_{browser} 为用户浏览行为评分,RF_{browser} 为浏览行为的奖励因子,具体公式如下:

$$R_{\text{browser}}=\lambda\times r_{\text{browser}}\times RF_{\text{browser}}.\tag{2}$$

(2)定义 R_{collect} 为用户收藏行为评分,RF_{collect} 为收藏行为的奖励因子,具体公式如下:

$$R_{\text{collect}}=\lambda\times r_{\text{collect}}\times RF_{\text{collect}}. \tag{3}$$

(3)定义 R_{share} 为用户分享行为评分,RF_{share} 为分享行为的奖励因子,具体公式如下:

$$R_{\text{share}}=\lambda\times r_{\text{share}}\times RF_{\text{share}}. \tag{4}$$

(4)定义 R_{comment} 为用户评论行为评分,RF_{comment} 为评论行为的奖励因子,具体公式如下:

$$R_{\text{comment}}=\lambda\times r_{\text{comment}}\times RF_{\text{comment}}, \tag{5}$$

其中,$\lambda=1$ 时表示用户发生该行为,$\lambda=0$ 则表示该行为未发生.奖励因子和用户行为评分 R_{jk} 公式如下所示:

$$RF_{\text{browser}}+RF_{\text{collect}}+RF_{\text{share}}+RF_{\text{comment}}=1,R_{jk}=R_{\text{browser}}+R_{\text{collect}}+R_{\text{share}}+R_{\text{comment}}. \tag{6}$$

记 u_j 为第 j 个用户,i_k 为第 k 个资源,$r_{j,k}$ 为用户 j 对资源 k 的评分,取值范围 $r_{j,k}\in[0,5]$.用户行为评分矩阵如表 2 所示.

表 2 用户行为评分矩阵

Tab. 2 User behavior scoring matrix

用户资源	i_1	i_2	…	i_k	…	i_{n-1}	i_n
u_1	$r_{1,1}$	$r_{1,2}$	…	$r_{1,k}$	…	$r_{1,n-1}$	$r_{1,n}$
⋮		⋮		⋮		⋮	⋮
u_j	$r_{j,1}$	$r_{j,2}$	…	$r_{j,k}$	…	$r_{j,n-1}$	$r_{j,n}$
⋮		⋮		⋮		⋮	⋮
u_m	$r_{m,1}$	$r_{m,2}$	…	$r_{m,k}$	…	$r_{m,n-1}$	$r_{m,n}$

综上所述,通过融合行业资源特征和用户行为特征两个维度的特征,构建体现行业特征的多维度用户画像模型.基于此脉络,画像构建模型分为 3 部分:特征标签提取、多维画像构建、画像用户分类与识别,构建“环评云助手”多维用户画像模型,如图 1 所示.

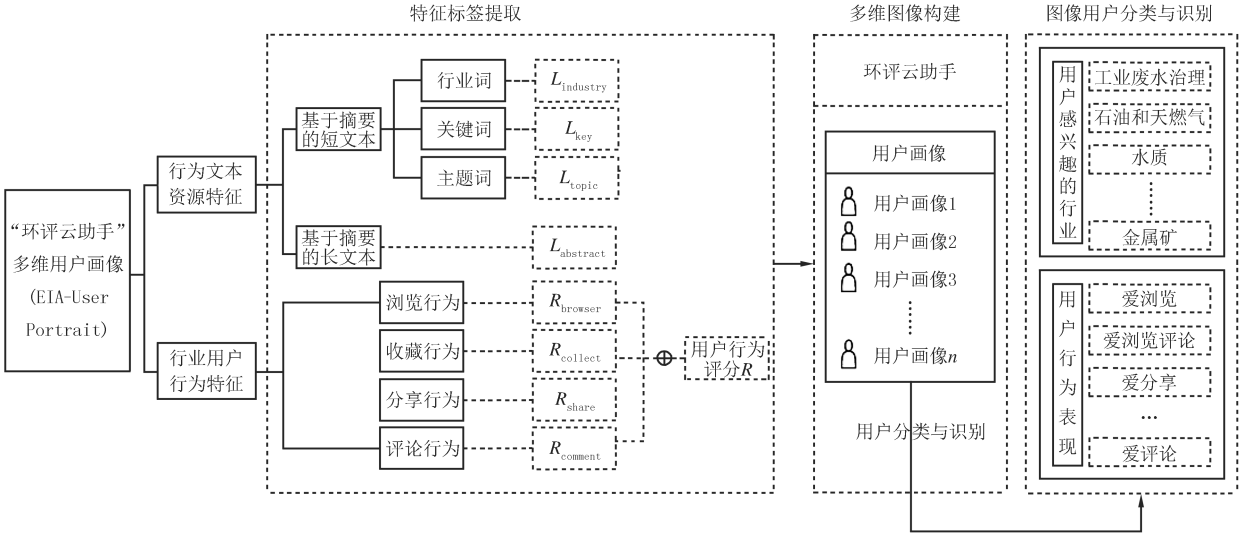

图1 “环评云助手”多维用户画像模型

Fig. 1 〝Eia Cloud〞 multi-user portrait model

在特征标签提取部分,将文本资源分为基于标题的短文本和基于摘要的长文本,进一步从标题短文本中提取行业词、关键词和主题词标签,从摘要长文本中提取综合摘要标签;又将用户行为分为浏览、收藏、分享和评论 4 项,根据自定义规则进行用户行为评分,最终将文本资源标签和用户行为评分合并设定为资源特征标签.根据提取的特征标签作为“环评云助手”多维用户画像标签,从而构建用户画像.并根据画像分析和总结对用户进行分类和识别,主要从“用户感兴趣的方面”“用户行为表现”两方面识别和描述用户.例如“一个爱分享对污水处理方面感兴趣的用户”、“一个爱评论收藏的金属矿开采行业的用户”等.

3　基于 DeepFM 的资源点击率(CTR)预测模型

本文的主要任务是给用户推荐其可能感兴趣的行业文本资源，因此需要将用户兴趣与资源信息相关联，从而进行建模.在第 2 节中，已经将用户感兴趣的资源信息和用户对此资源产生的行为数据进行语义提取以及构建评分矩阵，生成标签和用户画像模型.因此，将用户画像标签作为 DeepFM 的输入数据.

3.1　特征表示

由于用户画像标签的数据量大且属性种类繁多，使用 one-hot 编码后，数据维度高且稀疏.单个特征表达能力弱、特征组合数据量爆炸、分布不均匀会导致受训程度不均匀，所以需要通过 embedding 层将高维稀疏特征转化为低维稠密特征.但数据维度过高时，传入 embedding 层依旧会导致数据量爆炸，出现参数过多的情况.于是先引入 field 概念，可以将同一个特征经过 one-hot 编码生成的数值特征放到同一个 field，再将不同 filed 传入 embedding 层.尽管不同 field 的输入维度不同，但是 embedding 之后向量的维度均相同[12]，为模型后续 FM layer 和 DNN layer 的输入打下基础.本文与画像结合的特征表示结构如图 2 所示.

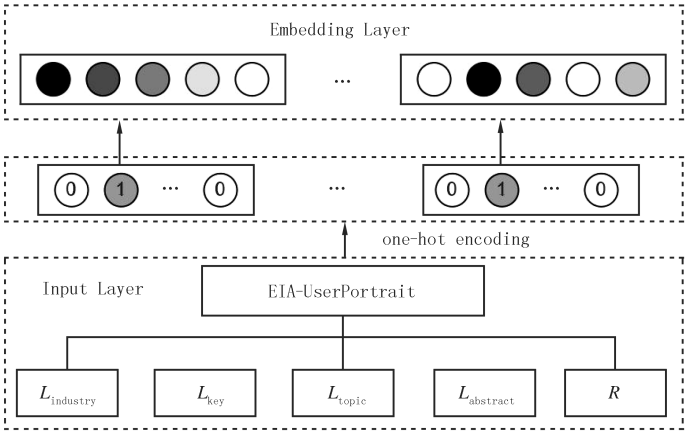

图2　特征表示结构图

Fig.2　Features structure

3.2　DeepFM 模型

DeepFM 是一种基于因子分解机的神经网络，其目的是学习低阶特征和高阶特征的交互.因此 DeepFM 由两部分组成，分别是因子分解机 FM 和深度神经网络(Deep Neural Network，DNN)，这两个部分共享相同的输入.本文将用户画像与 DeepFM 模型结合，其结构如图 3 所示.

DeepFM 模型公式为：

$$y' = \mathrm{sigmoid}(y_{\mathrm{FM}}, y_{\mathrm{DNN}}), \tag{7}$$

其中，$y' \in (0,1)$，y_{FM}是 FM 部分的输出，y_{DNN}是深度神经网络部分的输出.

FM 部分能用于学习特征之间的交互，每一个特征可以通过与其潜在的特征向量进行内积运算，来衡量其相关性.因此，FM 可以更好地学习数据中从未出现或很少出现的特征交互，有效地解决了本文行业资源特征和用户行为特征因数据稀疏而导致的特征交互难以表示的问题.FM 模型可以表示为：

$$y_{\mathrm{FM}} = \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < V_i V_j > x_i x_j, \tag{8}$$

其中，w_i 是特征 x_i 的权重，V_i 和 V_j 分别为特征 x_i 和 x_j 的潜在特征向量.

Deep Layer 部分是一个前馈神经网络，用于学习高阶特征交互.由于用户画像标签中特征输入向量为分类连续混合，具有高度稀疏、数据维度高等特点，经过 one-hot 编码后，神经网络的学习困难，学习效果不佳.因此需要在第一个隐藏层之前加一层 embedding 层，将长度不同的输入向量压缩为长度固定、低维、稠密的向量，再输入全连接网络层.同时使用 embedding 层可以使 FM Layer 部分和 Deep Layer 部分共享 embedding 输入层，使模型从原始特征中学习低阶和高阶特征交互.DNN 部分最终的输出结果为：

$$y_{\mathrm{DNN}} = \mathrm{sigmoid}(W^{|H|+1} a^{|H|+1} + b^{|H|+1}), \tag{9}$$

其中，$a^0 = [e_1, e_2, \cdots, e_m]$($m$ 为 filed 数量)作为 DNN 的输入，sigmoid 是激活函数，a^l、W^l、b^l 分别是第 l 层的输出、模型权重和偏差，$|H|$ 为隐藏层数.

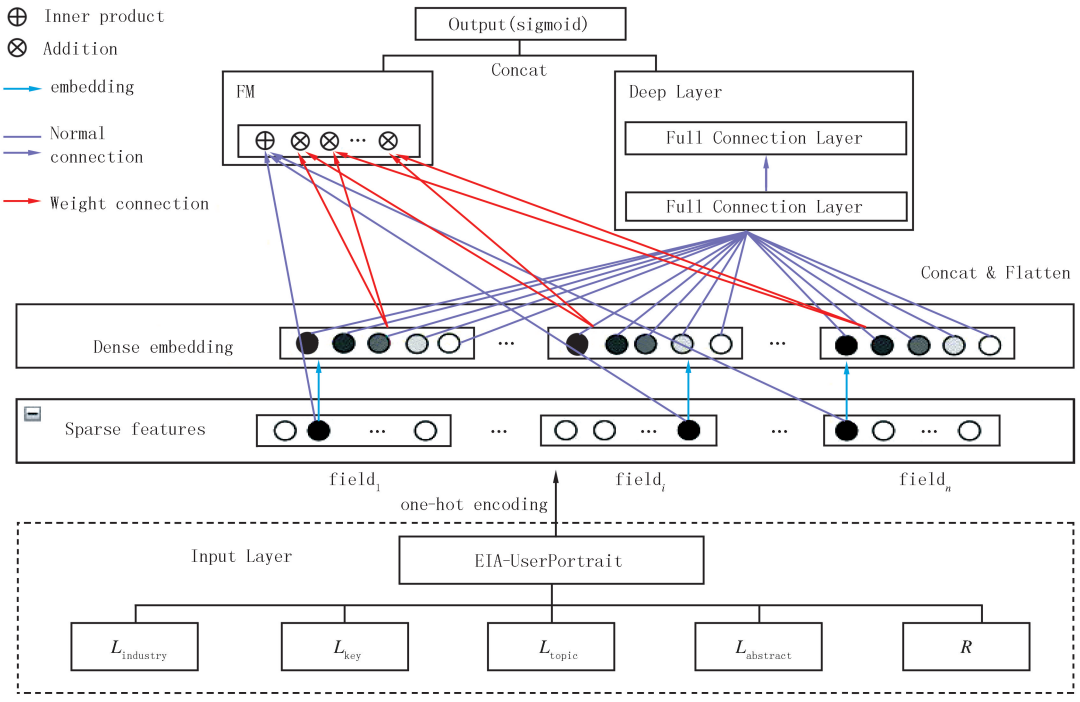

图3 DeepFM模型结构图

Fig.3 DeepFM structure

4 实验过程与分析

4.1 数据采集及预处理

本文筛选出“环评云助手”2019 年 11 月到 2021 年 1 月期间,2 119 名用户对中华人民共和国生态环境部发表的有关环境影响评价的 1 702 篇文章产生的 21 102 条数据,其中文章类型包括技术导则、技术规范、监测规范及相关行业标准等,行为数据包括浏览、收藏、评论及分享等.“环评云助手”APP 文本资源和用户行为数据均为未公开数据集,使用权限已由北京尚云环境有限公司授权,可作为论文数据集发表在期刊上.

数据预处理主要包括:过滤数据集中的空数据,根据哈尔滨工业大学实验室提出的停用词表,使用 jieba 分词库对数据集进行分词,并去除停用词、特殊符号和无意义的词等.

在上述数据集的基础上,进一步划分为资源信息数据集(Resource Information)和环评多维画像数据集(EIA-UserPortrait),数据集具体属性如下所示:

EIA-UserPortrait=(industry,key,topic,abstract,R),

Resource Information=(fileName,abstract,classification,flglml,gmjjdm),

其中,fileName 为资源名称,abstract 为资源摘要,classification 为资源类型,flglml 为分类管理名录,gmjjdm 为国民经济代码.

4.2 评价指标

本文实验以 AUC 和 LogLoss 为评价指标 0.

AUC(Area Under Curve)为受试者操作曲线(Receiver operating characteristic,ROC)下与坐标轴围成的面积,是衡量二分类模型优劣的一种评价指标.CTR 资源点击率预测任务作为二分类模型任务,研究表明 AUC 作为一个评价二分类问题广泛使用的指标,可作为评价其 CTR 预测性能的良好评价标准.LogLoss 是二分类模型的评价标准,其基于概率度量,用来表示预测值与真实值之间的差距.蒋兴渝等[15],GUO 等[12]和 LIAN 等[18]表示,对于 CTR 预测算法, AUC 提高 1‰也具有意义,因为推荐算法一般用于公司用户群体之间的推荐,如果用户数量非常大,它为公司收入增幅也自然会很大.

最后将整个数据集按 4∶1 的比例分割成训练集和测试集,并保证正负样本比例接近 1∶1.表 3 列出了

数据集的详细划分情况.

表 3　实验数据集统计表

Tab. 3　Experimental data-set statistics table

数据集划分	统计信息	数据量	数据集划分	统计信息	数据量
训练集	正样本	8 553	测试集	正样本	2 061
	负样本	8 449		负样本	2 039

4.3　实验结果与分析

实验分析主要包括如下内容:

(1)通过多次实验结果的比对,确定 LDA 主题模型的最优主题数目;

(2)基于相同参数,使用 DeepFM 模型分别对 Resource Information 数据集和 EIA-UserPortrait 数据集进行实验,测试多维用户画像对 CTR 预测模型的性能改进情况.与其他 CTR 预测模型作实验对比,通过比对实验结果,证明本文模型的有效性和优势.

4.3.1　LDA 最优主题数对比实验

为确定使 LDA 算法达到最优性能评价指标所对应的主题数,遍历了 1 至 51 之间 LDA 主题数目,每次增加的步长为 5,共 9 组实验.分别统计每组实验的困惑度值 $P(D)$,困惑度公式如下:

$$P(D)=\exp\{-\frac{\sum_{d=1}^{M}\ln\ p(w_d)}{\sum_{d=1}^{M}N_d}\},\tag{10}$$

其中,D 表示语料库中的数据集,共 M 篇文档,N_d 表示每篇文档 D 中的单词数,w_d 表示文档 d 中的词,$p(w_d)$即文档中词 w_d 产生的概率.实验结果如图 4 所示.

从结果可以看出,LDA 主题数目为 41 时,困惑度值最小,性能综合评价最好.

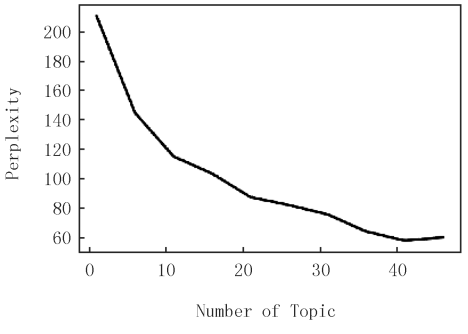

图4　LDA主题困惑度

Tab. 4　LDA Perplexity

4.3.2　与其他 CTR 预测模型对比实验与分析

为了验证所提模型的有效性,本文从以下 2 个类别中选择基线:(1)基于 Resource Information 数据集的 DeepFM 模型(R-DeepFM),(2)基于 EIA-UserPortrait 数据集的 DeepFM 模型(EUP-DeepFM).

实验还将基线对比模型分为两个部分:浅基线模型和深基线模型.浅基线模型实验使用 Resource Information 数据集作为各 CTR 模型的输入,深基线模型实验使用 EIA-UserPortrait 数据集,测试各 CTR 模型与用户画像结合的模型性能.

本文的浅基线模型为 R-(GBDT+LR)、R-FM、R-FNN、R-PNN 和 R-DeepFM,深基线模型是各 CTR 模型和用户画像的结合,即 EUP-(GBDT+LR)、EUP-FM、EUP-FNN、EUP-PNN 和 EUP-DeepFM.

表 4 展示了浅基线模型在资源信息数据集上的 AUC 和 LogLoss 结果,DeepFM 为本文 CTR 预测任务中使用的浅基线模型,观察实验结果可以看出 R-DeepFM 的性能均优于其他浅基线模型,因此本文 CTR 预测部分使用 DeepFM 模型.

为了进一步提升模型性能,将用户画像与各 CTR 预测模型结合,组成深基线模型,实验性能对比结果如表 5 所示.通过观察浅基线组与深基线组的模型性能比较可以看出,与用户画像模型结合在一定程度上提升了挖掘用户潜在兴趣的能力,使得 CTR 预测任务更加准确.在与其他 CTR 预测模型比较中,EUP-DeepFM 在 AUC 和 LogLoss 两方面的综合表现优于其他 CTR 预测模型,这说明本文提出的模型相比其他模型具有优势,也体现了用户画像和 DeepFM 模型的结合可以挖掘出更多有潜在价值的信息.

而且,基于"环评云助手"数据集进行实验时,EUP-DeepFM 模型比 R-DeepFM 模型在 AUC 值上提升

了 0.47%,LogLoss 值降低了 1.63%.EUP-DeepFM 模型的 AUC 值越接近 1 并且 LogLoss 损失值更低,说明该模型真实性更高,模型的预测性能更好,意味着更好的 CTR 预测和模型性能.其原因在于用户画像标签能挖掘出隐藏在用户和资源数据中潜在的信息,可以使二分类模型任务具有更高的预测准确率,为用户提供更加精准的、有效的个性化服务.

表 4 浅基线模型实验性能对比

Tab. 4 Comparison of experiment of shallow baseline model

模型	AUC	LogLoss
R-(GBDT+LR)	0.673 6	0.167 7
R-FM	0.861 0	0.191 4
R-FNN	0.952 4	0.223 2
R-PNN	0.952 6	0.191 7
R-DeepFM	0.967 8	0.176 2

表 5 深基线模型实验性能对比

Tab. 5 Comparison of experiment of deep baseline model

模型	AUC	LogLoss
EUP-(GBDT+LR)	0.781 1	0.159 7
EUP-FM	0.925 8	0.162 9
EUP-FNN	0.959 2	0.184 6
EUP-PNN	0.954 5	0.156 7
EUP-DeepFM	0.972 5	0.159 9

5 结 论

本文为“环评云助手”APP 构建行业用户画像和个性化推荐的研究工作提供了新的思路,部分解决了大数据时代 APP 中“信息过载”问题,为分析海量文本信息和精准找到信息提供了新的方法.针对“环评云助手”APP 中行业资源文本特征利用不充分、资源推荐精准较低的问题,提出了结合用户画像与 DeepFM 模型结合的推荐算法,更充分利用了环评行业文本资源特征和行业用户的行为特征,提升了推荐算法中 CTR 点击率预测率问题.实验结果表明,本文提出的模型有效提高了 APP 资源推荐的性能,具有一定的应用价值.

本文虽对“环评云助手”资源推荐存在的问题进行了研究,但本文提出的模型也存在一定的不足.本文使用的数据为用户历史数据,模型暂时没有考虑用户兴趣等特征随时间推移产生的变化.因此,在后续的研究工作中将进一步考虑用户的兴趣变化对模型的影响.

参 考 文 献

[1] TEIXEIRA C,PINTO J,MARTINS J.User profiles in organizational environments[J].Campus-Wide Information Systems,2008,25(3):329-332.

[2] QUINTANA R M,HALEY S R,LEVICK A,et al.The Persona party:using personas to design for learning at scale[C]//Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.New York:ACM,2017:933-941.

[3] 刘海鸥,黄文娜,姚苏梅,等.基于深度学习的移动图书馆用户画像情境化推荐[J].图书馆学研究,2019(21):57-64.
LIU H O,HUANG W N,YAO S M,et al.Situation recommendation for mobile library users' portrait based on deep learning[J].Research on Library Science,2019(21):57-64.

[4] 赵辉,化柏林,何鸿魏.科技情报用户画像标签生成与推荐[J].情报学报,2020,39(11):1214-1222.
ZHAO H,HUA B L,HE H W.User profile tag generation and information recommendations for science and tencnology intelligence[J].Journal of the China Society for Scientific and Technical Information,2020,39(11):1214-1222.

[5] GOLDBERG D,NICHOLS D,OKI B M,et al.Using collaborative filtering to weave an information tapestry[J].Communications of the ACM,1992,35(12):61-70.

[6] 张亮.基于 LDA 主题模型的标签推荐方法研究[J].现代情报,2016,36(2):53-56.
ZHANG L.Research on tagging recommendation method based on LDA topic model[J].Journal of Modern Information,2016,36(2):53-56.

[7] 熊回香,窦燕.基于 LDA 主题模型的标签混合推荐研究[J].图书情报工作,2018,62(3):104-113.
XIONG H X,DOU Y.Research on tag hybrid recommendation based on LDA topic model[J].Library and Information Service,2018,62(3):104-113.

[8] 熊回香,杨雪萍,高连花.基于用户兴趣主题模型的个性化推荐研究[J].情报学报,2017,36(9):916-929.
XIONG H X,YANG X P,GAO L H.Personalized recommendation research based on user interest topic model[J].Journal of the China Society for Scientific and Technical Information,2017,36(9):916-929.

[9] 熊回香,杨雪萍.社会化标注系统中的个性化信息推荐研究[J].情报学报,2016,35(5):549-560.
XIONG H X,YANG X P.Personalized information recommendation research based on combined condition in folksonomies[J].Journal of the China Society for Scientific and Technical Information,2016,35(5):549-560.

[10] 李兴华，陈冬林，杨爱民，等.基于用户兴趣-标签的混合推荐方法研究[J].情报学报，2015，34(5)：466-470.
LI X H，CHEN D L，YANG A M，et al.A study of mixed recommendation method based on user interest-tag[J].Journal of the China Society for Scientific and Technical Information，2015，34(5)：466-470.

[11] JUAN Y，ZHUANG Y，CHIN W S，et al.Field-aware factorization machines for CTR prediction[C]//Proceedings of the 10th ACM Conference on Recommender Systems.New York：ACM，2016：43-50.

[12] GUO H F，TANG R M，YE Y M，et al.DeepFM：a factorization-machine based neural network for CTR prediction[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence.New York：ACM，2017：1725-1731.

[13] 李维，闫晓东，解晓庆.基于改进 TextRank 的藏文抽取式摘要生成[J].中文信息学报，2020，34(9)：36-43.
LI W，YAN X D，XIE X Q.An improved TextRank for Tibetan summarization[J].Journal of Chinese Information Processing，2020，34(9)：36-43.

[14] 车蕾，杨小平.多特征融合文本聚类的新闻话题发现模型[J].国防科技大学学报，2017，39(3)：85-90.
CHE L，YANG X P.News topic discovery model of multi feature fusion text clustering[J].Journal of National University of Defense Technology，2017，39(3)：85-90.

[15] 蒋兴渝，黄贤英，陈雨晶，等.特征重要性动态提取的广告点击率预测模型[J].小型微型计算机系统，2022，43(5)：976-984.
JIANG X Y，HUANG X Y，CHEN Y J，et al.Advertising click-through rate prediction model based on dynamic extraction of feature importance[J].Journal of Chinese Computer Systems，2022，43(5)：976-984.

[16] 付芬，豆育升，韩鹏，等.基于隐式评分和相似度传递的学习资源推荐[J].计算机应用研究，2017，34(12)：3725-3729.
FU F，DOU Y S，HAN P，et al.Learning resource recommendation based on implicit scoring and similarity propagation[J].Application Research of Computers，2017，34(12)：3725-3729.

[17] 顾寰，杨长春，吴云，等.融合社区结构和个人兴趣的协同过滤推荐算法[J].计算机工程与设计，2018，39(11)：3420-3424.
GU H，YANG C C，WU Y，et al.Collaborative filtering recommendation algorithm combining community structure and personal interests[J].Computer Engineering and Design，2018，39(11)：3420-3424.

[18] LIAN J X，ZHOU X H，ZHANG F Z，et al.xDeepFM：combining explicit and implicit feature interactions for recommender systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.New York：ACM，2018：1754-1763.

Research on recommendation of "EIA Cloud" based on multidimensional user portrait and DeepFM

Li Tianyu[1]，Che Lei[1]，Ding Feng[2]，Tan Yue[1]

(1. School of Information Management，Beijing Information Science and Technology University，Beijing 100192，China；
2. Beijing Shangyun Co.，LTD.，Beijing 102208，China)

Abstract：“EIA Cloud” is an APP that serves users in the Environmental Impact Assessment. In view of the problems such as the surge of information，the insufficient use of text features and the low accuracy of real-time resource recommendation by users，the paper proposes a multi-dimensional user portrait model based on DeepFM combined with industry resources and user behavior to achieve CTR prediction. Firstly，the industry resource semantics is extracted，and then the user behavior is scored to build a multi-dimensional user portrait model. Finally，the model is applied to DeepFM to perform CTR prediction and achieve personalized recommendation with industry feature. Experimental data are obtained from “EIA Cloud”. The experimental results show that the model can effectively improve the AUC value of CTR prediction tasks and reduce the LogLoss value，which has certain application value.

Keywords：Environmental Impact Assessment；user portrait；tag generation；recommendation algorithm；DeepFM；CTR prediction

[责任编校　陈留院　赵晓华]