

基于数据挖掘技术的个性化学习模式研究

王仡捷

(安徽开放大学 信息技术与网络管理中心, 合肥 230022)

摘要:为了根据学生的学习情况和能力水平提供可定制的学习体验,提出了一种基于数据挖掘技术的个性化学习模式,该模式采用协同过滤算法对学生的学习行为进行分词,并为其推荐适合的学习资源和学习方式。实验验证表明,该模式可以有效地提高学生的学习效果和兴趣,同时也有助于教师更好地了解学生的学习情况和需求。

关键词:数据挖掘;个性化学习;协同过滤算法

中图分类号:TP311.13;G642

文献标识码:A

文章编号:2097-0625(2023)03-0092-05

一、引言

随着互联网技术和智能化设备的发展,网络学习逐渐成为一种受欢迎的学习方式。网络学习不仅有时间灵活、地域无限等优势,还可以让学生根据自己的兴趣爱好和学习习惯,选择适合自己的学习资源和方式^[1]。然而,在传统的网络学习中,学习资源和方式往往是通用的,缺乏对不同学生的个性化定制。如何利用数据挖掘技术,为学生提供个性化的、高效的学习模式,是当前亟待解决的问题^[2]。本文针对上述问题,提出了一种基于数据挖掘技术的个性化学习模式,能够根据学生的学习情况和能力水平,为其量身定制学习方案和资源。具体而言,在该模式中,使用协同过滤算法预测学生的学习行为,并根据学生的兴趣、能力等因素,推荐适合的学习资源和方式。通过实验验证,我们发现该模式可以显著提高学生的学习效果和兴趣,同时对教师和学生需求也有很好的反馈作用。

国外已经有很多的教育机构开发了相关平台,如智能教育系统(Knewton)、学习分析系统(Moodle)、以及对话式学习机器人(Socrates)。智能教育系统是利用人工智能技术设计的一个个性化的数字教学平台,能够根据学生的学习能力和兴趣等因素定制课程模块,满足学生的个性化需求。研究显示,在智能

教育系统支持下,学生的学习效果有了显著提高,且学生的学习积极性和满意度也获得了提高。Intayoad 等学者通过分析学生的行为,比如学习时间、学习履历、交互等行为以及学生成绩、兴趣爱好,从而了解学生的学习状态和需求,并为他们提供适合的学习方案和资源。这种方法能够有效地提高学生知识的消化吸收和理解,从而提高学习效果^[3]。

在科技的支持下,我国开始尝试性的推进个性化学习。比如利用大数据分析等高级技术来分析学生的需求 and 能力,为他们制定适合的学习计划。国内一些机构通过分析学生的性格特点,设计出以学生为主导、强调个体差异和情景变异的个性化学习方案,根据不同类别学生的不同特征,制定有效的个性化策略,极大地促进了个性化学习的效果,如 2020 年,朱佳等人针对国内的个性化学习方式进行了研究,提出一种基于数据挖掘的个性化学习方法^[4];张蓉等人提出一种适应用户兴趣变化的个性化学习算法,以此来解决传统学习不能反映用户特征的问题^[5];2019 年,彭红超等人提出一种基于协同过滤的个性化学习算法,从而解决提高课程推荐的准确性问题^[6],赵学孔认为可以从用户的个性化学习行为中获取用户评分,从而通过协同推荐算法来进行个性化推荐,并提高学习效率^[7]。

收稿日期:2023-04-03

基金项目:安徽省教育厅 2020 年度安徽高校自然科学基金项目(项目编号:KJ2020A1218)

作者简介:王仡捷(1981—),男,上海人,工程师,硕士。研究方向:计算机网络、数据挖掘。

二、个性化学习模式设计

(一) 协同过滤算法介绍

协同过滤算法是一种基于用户行为数据的推荐系统算法。该算法通过分析用户历史行为极其相似用户的行为,来推荐用户可能感兴趣的物品或信息^[8]。协同过滤算法可以分为两种类型:基于用户的协同过滤和基于项目的协同过滤。基于用户的协同过滤算法是找到与目标用户兴趣相似的其他用户来进行推荐。而基于项目的协同过滤算法则是对目标项目及其相关项目之间的相似度进行计算来推荐。在实际应用中,协同过滤算法需要处理海量数据^[9]。为了解决这个问题,许多高效的协同过滤算法被提出,如K近邻算法、SVD算法、潜在因子模型算法等。协同过滤算法在电商、社交媒体以及在线广告等领域得到了广泛的应用^[10]。例如,在电商平台上,协同过滤算法可以根据用户历史购买数据,为其推荐类似的产品;在社交媒体上,协同过滤算法可以推荐与用户兴趣相关的内容。此外,在线广告推荐中,协同过滤算法也能够根据用户历史点击记录和广告内容进行推荐,有助于提高广告点击率和转化率。然而,协同过滤算法也存在一些缺陷。例如,当用户和项目数目庞大时,算法可能会出现冷启动问题;在用户兴趣发生改变或者新物品出现时,算法也可能出现推荐结果不准确的情况^[11]。因此,在实际应用中,需要结合其他推荐算法来综合考虑不同因素,提供更加精准的推荐服务。

(二) 收集用户数据

用户数据主要包括用户基本信息、具体特征、学习习惯和学习偏好等相关的数据。系统在构建用户兴趣模型以及推荐结果时主要参考收集的用户数据。因此,数据的收集是用户兴趣建模的基本条件,收集的数据内容以及质量是构建用户兴趣模型和计算推荐结果的关键。一般来说,用户数据收集方式有两种,一种是显式方式,另一种则是隐式方式。本文中设计的系统采用两种收集方式相结合来完成用户数据采集。

1. 显式方法获取

进行显式用户数据采集时采取以下两种方法:

第一种方法,系统给用户提供一些兴趣爱好选择项目,提示用户进行挑选;给用户提出一些学习习惯

相关的问题,提示用户进行回答。系统通过这种较为明显的的数据收集方法,对首次使用系统的用户进行收集从而获取用户原始的学习爱好数据。

第二种方法,用户对学习完成之后的课程进行评分和文字评价。系统通过收集用户评分以及课程评价内容构建用户兴趣模型,从而推算出用户的兴趣爱好。这需要用户在课程学习完成之后积极地对课程进行评价。通过这种方式,系统能够有效获取用户更为真实的兴趣爱好数据,从而提高用户兴趣建模的效率。

由这两种显式收集用户信息的方法可以得知,用户的参与是显式收集用户信息的核心。系统要鼓励用户参与用户数据的真实填写过程和参与对课程的具体评价过程,才能有效保证用户信息采集效果。

2. 隐式方法获取

隐式方式获取用户信息是指系统不对用户进行明显的干扰,而是通过对用户使用系统的过程进行跟踪,从而获取用户使用系统的信息。例如对用户系统中浏览搜索的具体内容、对于某一课程的点击次数、收藏的课程记录等这些用户系统使用行为历史记录进行收集,以这些收集的数据作为依据,对用户的具体学习兴趣进行分析。

综上所述,使用多种数据采集方式,才能够使系统的数据来源更具准确性和完整性。为保证系统采集数据的质量,在设计此系统时,同时采用两种方式进行用户数据采集,优化系统课程推荐效果。

(三) 建立模型

1. 用户模型的表示

本文通过用户—项目评分矩阵表示法构建一个实验数据模型,可以将推荐系统表示为用户—课程评分矩阵($U \times D$)。具体来说, $u_i \in U$ 为用户集合,代表所有系统的用户, $d_k \in D$ 为课程集合,表示系统中的所有课程,而 $r_{ik} \in (U \times D)$ 则表示用户对课程的具体评分。

2. 改进的空值填补法

当前课程数量规模巨大,用户提供的评分数量过低,相似类型的用户之间对于某一个课程的评分缺乏共同的评分内容,使得很难完成用户间相似度的计算。当前行业较为流行的解决办法是使用空值填补

法。空值填补法是通过设置一个缺省值,来填补某项内容的评分空缺,减少评分矩阵之间的空缺,有效解决矩阵稀疏问题。

参考其他研究结果,此次系统设计过程中将结合课程评分等级以及课程最大卖点为主要内容,实现系统推荐算法。课程卖点具体来说包含课程的类型、长度、学习时长等,并设置课程评分相似的前提下,使用改进后的空值填补法,以获取新的缺省值为目的。

用 $(U \times D)$ 代表评分矩阵,该评分包括三个方面:类型用 s 表示;学习频率用 p 表示;学习时长用 z 表示。相应地,类型评分的权重为 Q_s ,学习频率评分的权重为 Q_p ,学习时长在评分中的权重为 Q_z ,且这三个权重之和为 1。

设在评分矩阵中,与课程 s_l 类型相同的课程有 w 个,其评分为 (s_1, \dots, s_w) ;与学习频率相同的课程有 m 个,其评分为 (p_1, \dots, p_m) ;与学习时长相同的有 g 个,其评分为 (z_1, \dots, z_g) 。改进的空值填补法的公式如下:

$$r_{ik} = \frac{Q_s \sum_{l=1}^w s_l}{w} + \frac{Q_p \sum_{l=1}^m p_l}{m} + \frac{Q_z \sum_{l=1}^g z_l}{g} \quad (1)$$

3. 评分标准化

每一位用户对自己学习进行评价的准则和习惯不一样,哪怕不同用户对同一课程的认可度相似,内容也可能会有有一定的差别。在协同过滤推荐算法中,必须依据其他用户给出的具体评分才能对目标用户的学习喜好进行准确预测。如果系统使用不同的用户评分,直接进行参考和推算,对于系统推荐的结果会带来一定的消极影响。这是因为在现实中,学习习惯不同的用户对同一产品的评价有可能相同,对某些课程给出的评分大致相同;学习习惯相同的用户对同一产品的评价因为具体评分的习惯不一样,从而使得对同一产品的具体评分产生很大的差别。因此,在建立模型时,不能使用个人评分标准,应该基于通用的整体评分标准,从而建立一个通用的标准化机制。在本文构建的系统中,采用了均值中心化的方法进行评分标准化设计,公式如下:

$$\delta_{ik} = r_{ik} - \bar{r}_{ui} \quad (2)$$

其中 r_{ik} 表示用户对课程的具体评分, \bar{r}_{ui} 表示 r_{ik} 的平均值, δ_{ik} 表示上述两者的差值。

(四) 用户相似度计算

1. 引入时间效用的用户相似度计算

用户对学习对象的兴趣具有随着时间的改变而改变的可能性,人们更加钟爱自己较为熟悉的课程向其他人进行推荐。在此推荐系统中,用户对课程的具体评分体现了用户进行评价时对课程的感兴趣程度,随着时间的推移,用户对课程的感兴趣程度将逐渐降低,意味着一段时间后用户进行的评价效用会有所衰减,具体通过遗忘函数 $J(t)$ 表示,其形式如下:

$$J(t) = \frac{ae^b}{(t+t_0)^c} \quad (3)$$

公式中 a, b, c, t_0 为变量,各项数值都大于 0, t 代表评分时间与目前的时间差。考虑用户的兴趣会进行转移,而且这种转移是与人类的自然遗忘规律一致的,因此本文设计的初值为 $a = 20, b = 0.42, c = 0.0225, t_0 = 0.00255$ 。

本系统在进行相似度计算时,采用了皮尔森相关系数的计算方式,该方式主要是引入了遗忘函数的概念,其计算公式为:

$$\begin{aligned} \text{sim}(u_i, u_j) = & \frac{\sum_{d_k \in D_y} (\delta_{ik} J(t_{ik}) - \bar{\delta}_{uj})}{\sqrt{\sum_{d_k \in D_y} J(t_{ik}) (\delta_{ik} J(t_{ik}) - \bar{\delta}_{uj})^2 \sum_{d_k \in D_y} (\delta_{jk} J(t_{jk}) - \bar{\delta}_{uj})^2}} \end{aligned} \quad (4)$$

2. 用户间信任度的计算

在实际生活中,用户进行学习面临对课程的选择时,会提前寻找其他用户的学习经验,并对那些更具有客观性、权威性的经验进行借鉴。在推荐算法构建的课程推荐系统中,目标用户与其他用户并没有紧密联系,也不相识。但是在系统平台中,目标用户可以通过其他用户对同一课程的评价以及与其他用户之间产生的联系等,完成对其他用户信任度的判断和预测。此种信任度是在一种特殊的环境下,目标用户对其他用户行为的具体相信程度,此种程度能够使用某种算法进行量化。

我们将用户 u_i 对用户 u_j 的信任程度量化进行定义时,用 $T(u_i, u_j)$ 来表示 u_i 对 u_j 的具体信任程度。

系统在表现用户对用户的信任时,有概率方法和渐进方法两种。概率方法中由 0 代表用户可信,1 代

表用户不可信;而渐进方法则是对用户行为正向作用逐步提升的信任度的度量,即目标用户对此用户提供的信息信任度随着此用户的信息正确程度上具有多信任的状态,信任度的数值可以不同。如信任度可以表示为极度不信任、不信任、信任、非常信任。

系统利用数据集的特点,对两种算法的结果进行加权计算,从而最终计算出用户的信任度。首先假设用户 u_i 对用户 u_j 进行了关注,同时用户 u_j 并未对用户 u_i 进行关注。当系统设置 u_i 为目标用户时, u_i 对 u_j 的信任度 $t(u_i, u_j)$ 为 1;但当 u_j 为目标用户的时候, u_j 对 u_i 的信任度 $t(u_i, u_j)$ 为 0。

然后,用 $t'(u_j)$ 表示用户 ($u_j \in U$) 在社会网络中的影响力,参考有关文献,本文确定用户的社会影响力包含五个方面的因素,分别为:粉丝数 f 、点评数 d 、鲜花数 h 、贡献值 g 、社区等级 m 。根据公式(5)可以得到用户的影响力产生的信任度。

$$t'(u_j) = \frac{f_{u_j}}{N} + \frac{d_{u_j}}{\sum_{i=1}^N d_{u_j}} + \frac{h_{u_j}}{\sum_{i=1}^N h_{u_j}} + \frac{g_{u_j}}{\sum_{i=1}^N g_{u_j}} + \frac{m_{u_j}}{M} \quad (5)$$

其中: f_{u_j} 、 d_{u_j} 、 h_{u_j} 、 g_{u_j} 、 m_{u_j} 分别表示用户 $u_j \in U$ 的粉丝数、点评数、鲜花数、贡献值、社区等级, $N = |U|$ 表示用户总的数量, M 表示用户最高的社区等级。

但是上述计算得到的结果比较粗糙,得到的数值范围比较大,无法确定具体的用户信任度,因此还需要对上述结果进行标准化处理。在处理过程中,主要是采用了以下公式进行:

$$t''(u_j) = \frac{t'(u_j) - \min}{\max - \min} \quad (6)$$

(6)式中, \max 为用户信任度中的极大值,相应的, \min 为用户信任度中的极小值。

根据(6)式即可以得到在 $[0, 1]$ 范围内的用户信任度取值。得到该值后,需要计算其加权评价分,计算公式如下:

$$T(u_i, u_j) = \frac{t(u_i, u_j) + t''(u_j)}{2} \quad (7)$$

最后,还需要考虑用户评分随着时间推移而变化的情况,将计算得到的相似度值进行改进,得到最终的用户相似度。改进的公式如下:

$$\text{sim}(u_i, u_j) = \frac{\text{sim}'(u_i, u_j) + T(u_i, u_j)}{2} \quad (8)$$

(五) 用户偏好预测

通过以下几个步骤可以得到用户偏好的预测。

第一步:设目标用户 $u_i \in U$ 。首先对上一步计算得到的结果进行排序,将排序中的前 n 个用户选择出来,这 n 个用户作为邻居用户 U' , $n = |U'|$ 。

第二步:将这 n 个用户对课程的评分 δ_{jk} 作为目标用户的评分参考,以此来预测目标用户对课程的评分值 r'_{ik} 。

第三步:对上述得到的评分值进行加权计算,得到评分的平均值。具体的计算公式如(9):

$$r'_{ik} = \frac{\sum_{j=1}^n \delta_{jk} \times \text{sim}(u_i, u_j)}{n} + \bar{r}_{ik} \quad (9)$$

其中 r_{ik} 表示用户对课程的具体评分, \bar{r}_{ik} 表示 r_{ik} 的平均值。

三、结果分析

(一) 实验数据

该系统进行实验的数据来源是使用爬虫软件进行采集的某教学系统中的真实用户数据以及课程数据。在进行数据采集时,先选取某一组用户为种子节点,选取之后,使用滚雪球抽样的方式来抽取其他用户并采集这些用户的信息,从而得到实验需要的数据。此次实验过程中采集数据包含 100 位用户的具体信息,如背景信息、行为信息以及社会信息等,同时还包含相关的 2 000 门课程的具体信息。

(二) 评价标准

该系统能够度量预测用户评分的准确率,当平均绝对误差(MAE)较大时,系统预测准确率低,系统推荐准确度低,反之亦然。以下为 MAE 值公式:

$$\text{MAE} = \sqrt{\frac{1}{|T|} \sum_{(i,j) \in T} |r'_{ij} - r_{ij}|} \quad (10)$$

其中: r'_{ij} 为测试集 T 上预测得到的评分。

(三) 实验

本文使用传统的基于用户的协同过滤推荐算法作为基准算法进行对比实验,具体的实验结果如图 1 所示。

通过图 1 的算法比较可以发现,本文的算法 MAE 值一直都比传统的算法低,说明本文的系统具有很好的精确度。

本文还测试了不同 n 值,对系统推荐结果的影响,具体的实验结果如图 2 所示。

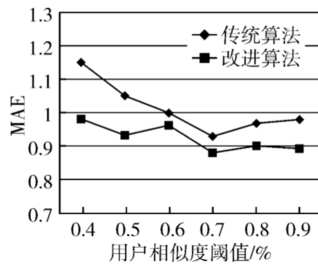
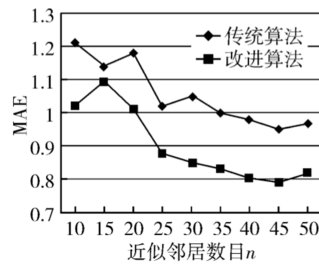


图 1 两种算法的 MAE 对比

图 2 不同 n 值的 MAE 对比

如图 2 所示,从算法验证的结果不难发现:无论 n 的取值为多少,改进后的算法都要比传统算法的 MAE 值低。这是因为在进行协同过滤算法改进时,充分将用户评分的时间衰减和用户信任度考虑进去,使得改

进后的算法在准确度方面有所提升。

四、结语

在传统的网络学习中,学习资源和方式往往是通用的,缺乏对不同学生的个性化定制。如何利用数据挖掘技术,为学生提供个性化的、高效的学习模式,是当前亟待解决的问题,为了解决上述问题,本文采用协同过滤算法对学生的学习行为进行分词,并为其推荐适合的学习资源和学习方式。通过实验验证,发现本文提出的算法具有很好的准确性,可以有效地提高学生的学习和兴趣,同时也有助于教师更好地了解学生的学习情况和需求。

参考文献:

- [1] 赵秀梅,赵宗昌,袁卫. 基于协同过滤的专业学习指导平台建设研究[J]. 计算机应用与软件, 2022, 6(14): 180-186.
- [2] 余学兵,黄沙,刘承启. 基于深度学习的网络资源优先协同过滤推荐[J]. 计算机仿真, 2022, 2(1): 12-14.
- [3] WACHARAWAN I, CHAYAPOL K, PUNNARUMOL T. Reinforcement Learning Based on Contextual Bandits for Personalized Online Learning Recommendation Systems[J]. Wireless Personal Communications, 2020, 115(4): 17-32.
- [4] 朱佳,张丽君,梁婉莹. 数据驱动下的个性化自适应学习研究综述[J]. 华南师范大学学报(自然科学版), 2020(4): 23-24.
- [5] 张蓉,刘志勇,李明星. 个性化学习路径研究综述[J]. 中小学电教, 2020(3): 55-58.
- [6] 彭红超,祝智庭. 人机协同决策支持的个性化适性学习策略探析[J]. 电化教育研究, 2019(2): 23-24.
- [7] 赵学孔,徐晓东,龙世荣. 协同推荐:一种个性化学习路径生成的新视角[J]. 中国远程教育(综合版), 2017(5): 34-38.
- [8] 陈超. 基于数据挖掘的个性化学习模式研究[J]. 电子设计工程, 2022, 12(1): 56-57.
- [9] 谢刚. 基于注意力机制与用户兴趣的序列推荐模型研究与实现[D]. 桂林: 桂林理工大学, 2022.
- [10] 石洋. 基于用户特征的酒店推荐模型研究[J]. 河南工程学院学报(自然科学版), 2022, 34(1): 34-39.
- [11] 张麦玲. 基于文本过滤技术的网络数据安全过滤方法[J]. 信息与电脑, 2022, 34(9): 60-62.

Personalized Learning Mode Based on Data Mining Technology

WANG Yijie

(Information Technology and Network Management Center, Anhui Open University, Hefei 230022, China)

Abstract: In order to provide customized learning experience according to students' learning situation and ability level, the personalized learning mode based on data mining technology is proposed. This mode uses the collaborative filtering algorithm to segment students' learning behavior and recommend suitable learning resources and learning methods for them. The experimental verification shows that this mode can effectively improve students' learning effect and interest, and also help teachers better understand students' learning situation and needs.

Keywords: data mining; personalized learning; collaborative filtering algorithm

[责任编辑 李潜生]