

基于用户画像的讲座信息 精准推送服务研究^{*}

杨传斌¹ 楼应凡²

(1. 浙江师范大学信息化办公室, 金华 321004; 2. 浙江师范大学教师教育学院, 金华 321004)

摘要: 目前高校图书馆的讲座推荐采用无差别推送方式, 导致学生在大量信息中无法及时掌握自己感兴趣的讲座信息, 且对讲座的效果也缺乏有效的反馈。用户画像作为大数据背景下个性化服务的一种设计工具, 为高校图书馆更精准地提供讲座推荐服务提供一种新的思路。本研究通过数字化校园平台收集学生基本信息、图书馆数据、学习活动数据、网络行为数据和学习行为数据, 按照基本信息、兴趣爱好、学术偏向三个维度构建用户画像, 采用滑动窗口机制实现画像权重的动态更新, 最后通过用户画像与讲座信息的匹配实现讲座信息的精准推送。

关键词: 用户画像; 讲座信息; 精准服务; 个性化推荐; 高校图书馆

中图分类号: G252 **DOI:** 10.3772/j.issn.1673-2286.2021.10.009

引文格式: 杨传斌, 楼应凡. 基于用户画像的讲座信息精准推送服务研究[J]. 数字图书馆论坛, 2021 (10): 60-65.

用户画像是指获取用户的专业背景、文化程度、兴趣爱好等与用户需求趋向相关的信息, 以此为基础进行模型化表示, 为用户制定特定标签^[1]。图情领域已有不少学者开展了将用户画像应用于图书馆精准服务方面的研究: 如刘海鸥等^[2]通过融合用户基本信息、内容偏好、互动、会话、情境标签的用户画像, 同时引入情景化推荐方法实现图书馆大数据知识个性化服务模式; 杨传斌等^[3]基于协同过滤推荐算法分析借书记录生成动态画像并考虑画像权重问题, 再根据用户画像匹配新书属性来实现图书馆的新书推荐功能; 吴智勤等^[4]收集用户属性、行为数据、社交网络大数据和互联网知识库并通过对社交网络分析来构建用户画像, 以此实现图书馆的个性化推荐服务、用户行为分析和知识问答功能。

学术讲座作为高校知识交流的平台, 可以增加大学生与优秀学者面对面交流的机会, 使得高校的学术氛围浓郁^[5]。高校图书馆通常将讲座推荐服务纳入业务范围, 向在校师生推送讲座信息。传统讲座推荐主要通过

宣传海报的方式, 但海报信息传播范围有限。随着互联网的高速发展, 通过QQ、微信或钉钉等软件群发消息, 或通过微信公众号进行推送是目前高校图书馆推送讲座信息的主要方式。它们能很好地将讲座信息及时送达学生, 但这种无差别推送也会使学生淹没在海量信息中, 以致错过感兴趣的讲座。如果充分利用用户画像技术来分析学生偏好、发现学生潜在需求, 再精准匹配相对应的讲座资源, 就能为学生个性化推荐讲座信息, 甚至可以根据学生标签信息, 有针对性地开展讲座活动。除此之外, 用户画像还会根据数据的动态更新而不断更新, 紧跟学生的兴趣变化趋势。笔者通过数字化校园平台收集学生数据, 构建用户画像, 最后通过用户画像与讲座信息的匹配实现讲座信息的精准推送。

1 用户画像构建

用户画像的构建主要包括数据资源层、数据处理层和用户画像层。在数据资源层利用高校各个系统和

^{*} 本研究得到教育部人文社会科学研究规划基金项目“智慧校园环境下图书馆用户画像及其应用研究”(编号: 18YJA870015)资助。

平台收集数据,在数据处理层对数据进行预处理,并从基本信息、兴趣爱好、学术偏向三个维度构建标签体系,在用户画像层生成画像并实现画像更新。

1.1 数据资源层

学生数据来自学校数据中心及各个有关应用系统

(如科研系统、教务系统、图书馆移动服务平台、校园网管理平台等),如表1所示,数据类型主要分为两类:一是静态数据,即变化较慢相对稳定的数据;二是动态数据,即实时变化更新的信息数据。其中静态数据包括学生的一些基本信息,动态数据主要包括学习活动数据、图书馆数据、网络行为数据和学习行为数据,画像更新的操作也主要是针对这一类数据。

表1 学生数据类型及来源

数据类型	分 类	属 性	数据来源
静态数据	学生基本信息	姓名、性别、学号、民族、籍贯、学位学历、研究方向、在籍情况、所属学院、所属专业、年级、学习经历、爱好特长、担任职务、必修课程、选修课程	研究生系统、教务系统
动态数据	学习活动数据	参加社团、学术报告、讲座、学校及学院活动、承担课题、科研成果	学工系统、科研系统 研究生系统
	图书馆数据	图书馆借阅书籍情况、数字图书馆访问情况	图书管理系统、 数字资源控制访问平台
	网络行为数据	访问网络的行为	防火墙日志系统
	学习行为数据	Blackboard、超星泛雅、Moodle等网络课程学习平台	网络学习平台

1.2 数据处理层

学生数据主要有结构化、半结构化和非结构化3种类型。结构化数据相对较容易收集,如学生基本信息、学习活动数据,这类数据只需要进行简单的清洗和挖掘就可以形成标签;但半结构化数据和非结构化数据相对来说需要更多时间进行前期的预处理,如网络行为数据和数字资源访问数据,其原始数据大部分存在缺少值、数据异常、冗余、噪声等问题,为了避免这些问题对用户画像带来混淆,所以要对数据进行提取、清洗、标准化3种方式操作进行预处理来保证数据的准确度和有效性。

1.2.1 网络行为数据处理

由于师生在校园内上网需要经过学校统一身份认证系统认证,这会在防火墙上留下网络访问日志,每一条日志内容包括访问时间、IP地址、上网账户信息、访问页面URL、访问数据库等内容。对网络访问日志的全面分析能获得学生的所有网络访问行为,但日志量非常庞大,自编程难以建立全面的日志分析系统,因此有必要借助日志管理分析工具软件(如日志易)来实现日志

的采集、存储、处理和分析,其中提供搜索和统计功能可以单项进行测算,提供的程序接口可以进行大量数据的处理。目前笔者所在单位的日志系统存储了约3年的网络行为日志,每天的日志数量在5 000万条以上,可以有效地开展基于日志的网络行为画像研究。

为了实现讲座的精准推荐服务,根据学生的网络访问行为总结其兴趣爱好构建用户画像是必要的。首先,需要对收集到的日志进行网站名归类,采用机器学习算法对数据集进行训练,根据分类好的训练集得到类效果最优的映射规则,建立网页分类器;然后,采用N-Gram方法来实现URL特征提取,N-Gram的基本思想是先将文本以N字节进行划分和滑动窗口操作,形成序列,计算所有字段出现频率;最后,按照阈值进行过滤。把经过处理的网络日志数据作为网站名分类的数据集,参考Alexa网站上对网站名的分类,采用面分类法对网站进行分类,每一分类下根据排名选取10~30个网站为子分类,最后将URL网站映射到对应的类别中(见表2)。经过试算,按此分类,能将约87%的网络行为划到十大类别内,随着挖掘学生浏览网站数量的增多,分类中的内容会越来越完善^[6]。

把每位学生的网络行为中最多的4~6个类别作为其兴趣爱好标签,若需要更加细化其兴趣爱好则需要

表2 URL数据源网站名分类

类 别	子类别
学习	出国留学网、古诗文网、有道词典、中国知网、道客巴巴、慕课网、菜鸟教程、新东方、超星、考研网、中国教师网、维普
视频	搜狐视频、爱奇艺、优酷、腾讯视频、豆瓣电影、哔哩哔哩、电影网、360影视、PPTV、电影天堂、发现频道、暴风影音、百度视频、芒果TV、乐视网、央视网、美剧天堂
游戏	4399游戏、7k7k游戏、虎牙直播、斗鱼TV、企鹅电竞、1717337游戏、九游手机网、游民星空、游侠网、181833手游网、游迅网、178游戏网
社交	QQ、微信、微博、百度贴吧、钉钉、知乎、豆瓣、天涯、虎扑、世纪佳缘、珍爱网
搜索	百度、搜狐、搜狗、360、必应
科技	中关村在线、苹果官方、站长之家、太平洋电脑、华军软件园、2345软件大全、博客园
阅读	起点中文、喜马拉雅、纵横中文网、潇湘书院、晋江文学、FM简书、笔趣阁、豆瓣
音乐	QQ音乐、网易云音乐、千千音乐、随心听、酷狗音乐、九酷音乐网、酷我音乐、音乐巴士
购物	淘宝、京东、天猫、当当网、华为商城、唯品会、苏宁易购、拼多多、亚马逊、聚划算
下载	百度网盘

单独列出某子类别作为标签,将其访问的次数作为权重,生成的标签是科普类讲座、非学科性讲座推荐的重要依据。

1.2.2 图书馆借阅数据的提取

通过图书管理系统提取学生借阅书籍的中图分类号,再根据中图分类号所代表的图书类别标记该学生的兴趣爱好,学生所借图书的数量和借期长短可用来计算对应的权重。为了验证该方法的可行性,通过图书管理系统采集某高校图书馆8个月内11 862名学生的110 104条借书数据,其中包含9 896种中图分类号,说明根据中图分类号来描述学生的兴趣爱好是相当细分的。若可选取中图分类号的前面部分,只选首字母则只有22个大类,如A为马列、B为哲学宗教、T为工业技术等,但用它去匹配讲座明显范围太广,为此将借书的中图分类号分为三个层次:第一为大类即为首字母,共22个标签;第二为中类即为首3个字母的,共约500个标签,如TP3为计算机技术;第三为细类,即完整的分类号,如TP311.5为软件工程,全部的标签约有10 000个。通过这样的层次划

分,学生借一本图书就形成3个标签,分别从三个层次去匹配讲座,兼顾了匹配成功率和精准度。

1.2.3 数字资源访问行为的提取

图书馆数字资源访问行为分析是了解学生学习和研究兴趣的重要来源,数字资源分为本地资源和外网资源两类,本地资源一般通过校园网直接访问,外网资源通过校园网出口IP地址验证访问,若在校园网外则通过VPN/WEBVPN访问,要取得其访问行为需要分析防火墙日志,且比前述的网络行为分类统计复杂许多。目前大多数高校图书馆都部署了统一检索系统,而统一检索系统可实现基于中外文资源元数据及多种文献类型的一站式检索并能直接定位到来源数据库获取全文;部分高校还部署了电子资源访问控制系统,通过身份认证进行访问控制,具备资源使用统计、读者行为分析等功能。为了掌握学生的学习研究兴趣点,通过这两个系统的日志能直接获得所有学生的检索词及其频度,计算后可作为标签的名称和权重,简化了学生研究兴趣的获取。

1.3 用户画像层

1.3.1 用户画像维度体系

根据数字化校园平台收集到数据的分析以及用户画像的需求,设定用户画像维度体系,由基本信息、兴趣爱好和学术偏向三个维度组成。基本信息维度包括姓名、性别、学号、民族、籍贯、学位学历、在籍情况和年级;兴趣爱好维度包括参加社团情况、参加学术报告情况、已参加讲座情况、参加学校学院活动情况、图书馆借阅情况、数字图书馆访问情况、网络课程选取情况、经常访问网站信息、爱好特长和担任职务;学术偏向维度包括研究方向、所属学院、所学专业、学习经历、必修课程情况、选修课程情况、承担课题情况。

1.3.2 标签生成

用户标签分静态标签和动态标签两类,静态标签是最基础的标签,由静态数据(如姓名、性别、民族、籍贯、学位学历、在籍情况、所属学院、所学专业等)形成。将这类静态标签直接写入用户画像,这类标签就是

用户画像的基础标签。对于学生的静态数据采用本体的建模方式, 构建定性用户画像, 画像产出形式为关键词法。动态标签是根据动态数据形成, 动态标签根据形成方式不同, 分为统计类动态标签和挖掘类动态标签。对于统计类动态数据, 根据建模方式的不同, 分为基础数据统计和网络日志数据统计。对于基础数据的统计, 这类数据主要包括网络课程数据等。采用融合用户兴趣的建模方法, 采用向量空间模型来进行表示, 即 $ID\{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \dots, (\alpha_n, \beta_n)\}$ 。其中, α_n 表示特征项, β_n 表示 α_n 在整个模型中的权重, 即用户对 α_n 的喜爱程度, 当 β_n 越大, 说明用户对这个事情越感兴趣, 确定当数值超过阈值 N , 则将它加入画像中。对于网络日志分类统计是在基础数据统计之上引入占比的概念, 根据网络日志数据的分析, 每一个学生都会生成十大类网站访问数据量, 可以将数据表示为: $Student\{(t_1, p_1, n_1), (t_2, p_2, n_2), \dots, (t_m, p_m, n_m)\}$ ($m < 11$), 其中, t_m 表示第 m 个网站主题, p_m 表示学生对第 m 个网站主题浏览量占总网站的比例, n_m 表示学生访问第 m 个网站主题的次数, 之后可以将占比排名前 N' (阈值 N'') 的类别标签加入画像。对于挖掘类动态数据, 第一部分是基础数据挖掘, 这类数据主要包括学生参与的各类活动以及参加过的讲座信息。对于这一类数据一般采用 TF-IDF 方法即词频-逆向文件频率计算方法, 将所得的动态标签标记到用户画像中。

1.3.3 用户画像的生成

在对数据进行清洗、集成、分析的基础之上生成标签, 将标签打在目标学生的画像模型上, 通常用一组标签来描述一个学生, 这些标签从多个维度不同层面描述学生整体特征。最终生成多层次画像标签, 主要包含 3 个方面内容, 即基本信息、兴趣爱好和学术偏向。形成用户画像之后, 可以据此实现对学生需求进行预测, 通过相似度计算来实现讲座信息的精准推荐服务。例如, 某同学的用户画像如下。

{基本信息: 女, 202020200871, 汉, 浙江绍兴, 研究生, 在籍, 研二;

兴趣爱好: 摄影协会, “基于核心素养的小学数学解题思想观”, “新自由主义情境下教师专业理论的跨国研究”, “语文作业与简单课堂”, 古典舞, 阅读, (科技类, 31.4%, 2 135), (音乐类, 20.5%, 1 243), (G434, 计算机化教学), (TP391.1, 文字信息处理);

学术偏向: 教师教育学院, 教育技术学, 电子化学学习资源与环境研究, 教学设计, 教育理论, 教育改革研究, 智慧教育}

1.3.4 用户画像的兴趣衰减和权重

随着时间的推移, 学生的兴趣也会不断变化, 为了更好地根据当前学生的状态来推荐讲座, 所以需要考虑学生最近新增加的行为或者爱好。对于静态数据权重保持不变, 动态数据将充分利用窗口滑动机制, 定时更新标签权重。每个用户画像动态数据标签都有相应的权重 W , 画像更新的间隔时间定为 T , 计算学生的动态标签权重变化, 根据变化情况进行增删, 并将更新后的标签数据存储到数据库中。

记学生前一次的标签权重为 W_{last} , 每间隔时间 T 后, 学生原有的兴趣程度会随着时间推移发生相应的减弱, 加入时间衰减因子 α , 新增加的标签权重记为 W_{new} , 充分考虑学生的兴趣随着时间发生变化的用户画像标签, 更新后的标签权重为 W_{update} , 计算过程为: $W_{update} = W_{last} \times \alpha + W_{new}$ 。关于时间衰减的函数, 根据发生时间的先后为用户行为数据分配权重。时间衰减是指用户的行为会随着时间减弱, 历史行为和当前的相关性不断减弱, 在建立与时间衰减相关的函数时, 套用牛顿冷却定律数学模型。其描述的场景是一个温度高的物体在温度低的环境下, 物体的温度要降低, 周围环境温度要上升, 最后二者的温度一致, 在这个平衡的过程中, 物体的温度 $F(t)$ 是随着时间 t 的增长而呈现指数型衰减, 其温度衰减公式为: $F(t) = \text{初始温度} \times \exp(-\alpha \times \text{间隔的时间})$, 其中 α 为衰减常数, 通过回归可计算得出。例如: 指定 45 分钟后物体温度为初始温度的 0.5, 即 $0.5 = 1 \times \exp(-\alpha \times 45)$, 求得 $\alpha = 0.0154$ 。应用到我们的兴趣爱好, 假设 1 个月后的兴趣爱好下降到原来的 90%, 即 0.9, 按照牛顿冷却定律 $0.9 = 1 \times \exp(-\alpha \times 1)$ 求得冷却系数 $\alpha = 0.1050$, 也可根据不同兴趣爱好特征设置不同的衰减常数 α 。具体实践中 T 可选 1 个月, 即每个月初计算一次学生的兴趣爱好。

2 讲座信息的精准推送服务

2.1 讲座特征提取

讲座推荐是用户兴趣需求和讲座匹配的过程, 因此

需要把讲座的属性特征做向量处理,以方便计算,处理方式与画像的生成方式类似。讲座的特征提取包括讲座名称、类别、主讲人基本信息、主讲人所属学校、主讲人专业方向、承办学院、主题词(3~10个),其中主题词生成的时候要考虑与构建的用户画像匹配,如根据讲座内容设置主题词为“科技类”、“音乐类”(与网络行为匹配)、“软件工程”(与借书记录TP311.5匹配)等。

2.2 推荐算法

利用大数据平台构建用户画像以及提取讲座的特征,再通过相关推荐算法找到讲座的目标群体,在学生群体里查找该类群体,将讲座信息内容发送给相关学生。比较现有的推荐算法的优缺点之后,笔者采用的是基于内容的推荐算法和基于物品的协同过滤算法^[7]。

2.2.1 基于内容的推荐算法

基于内容的推荐算法可以根据学生各种信息数据组成的画像来找到用户喜欢的讲座类型,通过对讲座的相关信息统一成一套标准化的描述统计方式,然后计算得到该讲座是否是某学生感兴趣的类型,如果是就推送给学生。实现过程采用基于知网(HowNet)的语义相似度^[8]。

将用户画像中的关键词和即将举办的讲座关键词,根据知网相似度进行对比,若有关键词相同则直接推送;若没有相同关键词,但有多个关键词语义相似度达到或超过阈值,也将讲座信息推送给学生;若达不到阈值,则不推送讲座信息。

2.2.2 基于物品的协同过滤算法

基于物品的协同过滤算法就是收集用户之前的一些喜爱物品数据,在此基础上为用户推荐类似的物品。本文选用Jaccard相似度来实现。当 $Jaccard(X, Y)$ 数值越大,说明讲座之间越相似^[9],当新旧讲座之间的相似度达到阈值,则将新讲座信息推送给参与过旧讲座的学生。

2.3 推荐结果的生成和发送

许多高校在建设智慧校园过程中都建立了基于统

一身份认证的消息中心,根据用户画像与讲座资源匹配的情况定时定向地将消息发送给学生,实现消息发送的自动化,减轻管理人员的负担。在消息通知中心可以选择邮件、短信、钉钉、微信和智能渠道的方式向学生推送合适的讲座信息。

在推送讲座信息下方有感兴趣/不感兴趣选项,及时了解学生对推送内容是否感兴趣;在推送中及时提醒学生去讲座报名系统报名。在实践中,可能会存在相似度阈值偏高或者学生时间冲突,导致在讲座报名系统中报名人数明显偏少,因此可以相应降低相似度阈值,再次向没有收到推送但在相似度范围内的学生发送讲座信息,及时增加报名人数,确保有基本的参加人数。

3 结语

高校讲座资源丰富,但如何在众多讲座资源中满足学生的个性化资源需求是当前讲座管理需要面对的重要挑战。随着大数据技术、用户画像技术等不断发展,为学生提供个性化服务也有了更大的发展空间。通过学生基本信息、兴趣爱好、学术偏向三个维度构建学生的多维画像模型,结合用户画像与讲座属性的匹配及推荐算法实现个性化推荐,讲座举办方只需提供讲座信息,其余都可以通过系统自动完成推荐,随着应用数据的不断深入完善和算法的不断优化,讲座信息的服务也将越来越精准。除给学生提供精准服务外,还可以通过分析用户画像中兴趣爱好维度和学术偏向维度中的标签变化来了解学生需求,对讲座开设的内容提出建议,有针对性地邀请有关专家开展讲座,提高讲座的质量和听众满意度。

参考文献

- [1] 曾建勋. 精准服务需要用户画像[J]. 数字图书馆论坛, 2017(12): 1.
- [2] 刘海鸥,姚苏梅,黄文娜,等. 基于用户画像的图书馆大数据知识服务情境化推荐[J]. 图书馆学研究, 2018(24): 32, 57-63.
- [3] 杨传斌,于冉,付敏峰. 智慧校园环境下图书馆新书推荐系统的设计与实现[J]. 高校图书馆工作, 2021, 41(4): 43-47.
- [4] 吴智勤,柳益君,李仁璞,等. 基于社交网络的高校图书馆用户画像构建研究[J]. 图书馆学研究, 2018(16): 25, 26-30.
- [5] 高海燕,王玲玲. 高校图书馆信息素养培训现状调查与分

- 析——以“双一流”高校图书馆专题讲座培训为例[J]. 图书馆工作与研究, 2019(2): 113-120.
- [6] 蒋丽. 大学生网络行为与学习成绩的相关性研究[D]. 金华: 浙江师范大学, 2020.
- [7] 王茹芳, 宁璐. 基于用户画像的图书馆推荐系统研究[J]. 图书馆建设, 2020(S1): 100-102.
- [8] 朱新华, 马润聪, 孙柳, 等. 基于知网与词林的词语语义相似度计算[J]. 中文信息学报, 2016, 30(4): 29-36.
- [9] 李晓敏, 熊回香, 杜瑾, 等. 智慧图书馆中基于用户画像的图书推荐研究[J]. 情报科学, 2021, 39(7): 15-22.

作者简介

杨传斌, 男, 1966年生, 硕士, 高级工程师, 研究方向: 校园大数据挖掘与知识发现、智慧教育和智慧图书馆。

楼应凡, 女, 1997年生, 硕士研究生, 通信作者, 研究方向: 校园大数据挖掘与知识发现、智慧教育和智慧图书馆, E-mail: louying_f@163.com。

Research on Precise Lecture Information Push Service Based on User Portrait

YANG ChuanBin¹ LOU YingFan²

(1. Information Office, Zhejiang Normal University, Jinhua 321004, China; 2. School of Teacher Education, Zhejiang Normal University, Jinhua 321004, China)

Abstract: At present, lectures in university libraries are recommended to be pushed in an undifferentiated way, which makes students unable to grasp the information of lectures they are interested in in time and lack effective feedback on the effects of lectures. As a design tool of personalized service under the background of big data, user profile provides a new idea for university library to provide lecture recommendation service more accurately. This study through the digital campus platform students basic information, library collection data, learning activity data, the network behavior and learning behavior data, according to the basic information, hobbies and academic to three dimensions to build user profiles, profiles weight is realized by using sliding window mechanism of dynamic update, the last of profiles by user and lecture information matching precision is recommended.

Keywords: User Portrait; Lecture Information; Precision Service; Personalized Recommendation; Academic Library

(收稿日期: 2021-09-18)

■ 书 讯 ■

《〈汉语主题词表〉构建研究》

《汉语主题词表》自1980年诞生以来,在我国情报检索语言发展历程中发挥了应有作用。作为国际上最大型的综合性主题词表,一方面体现了我国图书情报领域众多专家的集体智慧,带动了我国众多专业性汉语主题词表的编制与应用;另一方面提升了我国文献资源信息组织能力和水平,强化了中文文献的文档索引和查全查准率。2009年,中国科学技术信息研究所开始了《汉语主题词表》在网络环境下的改造与重修工作。《〈汉语主题词表〉构建研究》记录十年来新型《汉语主题词表》的探索经历与实践过程,主要表述汉语类主题词表的发展演变路径,记录《汉语主题词表》的历史作用与现实意义,展示新形势下《汉语主题词表》重构的框架、思路、流程和模式,归纳《汉语主题词表》构建中探索的标准规范、基础词库、协同编制平台和范畴体系,阐述概念建设、关系构建和范畴分类等核心构建业务,详述《汉语主题词表》服务系统。曾建勋等著的《〈汉语主题词表〉构建研究》,2020年由科学技术文献出版社出版。