基于 KD-Tree 聚类的社交用户画像建模

万家山 陈 蕾 吴锦华 高 超
(安徽信息工程学院大数据与人工智能学院 安徽 芜湖 241000)

摘 要 传统的信息推送服务普遍缺少对社交用户具体情况的考虑,存在推荐信息针对性不强、系统转化率低等问题。针对上述问题,提出了一种基于用户画像的智能信息推送方法。借助智慧学习平台的用户数据,主要通过 KD 树来实现在 KNN 聚类算法中分析用户偏好和行为特征,进而将用户进行类别划分。首先,通过分析聚类中心将每一类用户抽象成高度精炼的短文本,形成具有代表性的标签;其次,根据社交用户个体的标签权重值,结合业务需求进行二次建模来构建用户画像模型,进而逐步细化模型;最后,借助协同过滤推荐算法产生推荐。用户画像不仅提高了数据的可用性和价值,还使分析者从大量的用户数据中摆脱出来,快速地协助分析者做好精细化分类,达到了较好的推荐效果。

关键词 在线资源推送,KD 树,用户画像,个性化推荐
中图法分类号 TP311.1 **文献标识码** A

Persona Based Social User Modeling Using KD-Tree

WAN Jia-shan CHEN Lei WU Jin-hua GAO Chao
(School of Big Data and Artificial Intelligence, Anhui Institute of Information Technology, Wuhu, Anhui 241000, China)

Abstract Traditional information push service takes little consideration of specific needs of social network users in particular conditions, hence it has poorly-targeted recommendations and low-rated system transformation. Responding to these problems, this paper proposed an intelligent push method based on user personas. By analyzing user data of intelligent learning platforms KNN clustering algorithm realized by KD-Tree is used to analyze user preferences and behavior characteristics, and then classifies user categories. First, through clustering center analysis, each type of users is abstracted into a highly-refined short text to form a representative label. Second, on account of label weight value of individual users and different service demands, user personas are modeled two times for refinement. Finally, recommendations are made by collaborative filtering algorithm. User personas will enhance the usability and value of user data. In addition, they may free analysts from large volumes of data, and help make fine classifications and thus more accurate recommendations.

Keywords Online learning resource push, KD-Tree, User personas, Individualized recommendation

1 引言

随着互联网的发展产生了海量数据,如传统的 IT 时代,IT 系统围绕业务服务,在这个服务的过程中沉淀了很多数据,在线学习平台同购物、社交等其他类型的网站相似,普遍面临着信息过载的问题。对于用户或学习者而言,仅仅通过简单的检索功能从海量的视频条目中寻找自己感兴趣的视频,将会浪费大量时间。但是在 DT 时代就不一样了,数据是现实世界的虚拟化表现,数据本身构成了一个虚拟系统。对于视频提供商来说,如何利用这个虚拟系统的数据优势,准确高效地为用户或学习者推送优质的资源,提高用户黏性,从而实现虚拟系统的价值最大化成为了一个亟待解决的难题。在这样的背景下,如何使得在线学习平台的更加智能,科学的用户画像建立和智能的资源推送算法就显得十分重要。

在线学习平台的种类众多,针对不同类型的用户往往有不同甚至相冲突的需求,不可能做出一个满足所有用户需求的在线学习平台。因此,在线学习平台确定目标用户就显得至关重要。为了能够准确聚焦目标用户的动机和行为,最早由 Alan Cooper[1] 提出了 persona 的概念:"Personas are a concrete representation of target users." Persona 是真实用户的虚拟代表,是建立在一系列真实数据之上的目标用户模型,即用户画像的前身。已有学者就传统的协同过滤方法中矩阵的稀疏性带来的冷启动问题进行了研究,以标签的视角王卫平等[2]提出了基于标签和协同过滤的混合推荐方法,刘凯鹏等[3]利用标签来构建用户、资源模型以达到推荐排序的效果。之后又有学者提出,虽然标签的引入能够有效缓解传统协同过滤中的冷启动问题,但忽略了用户评分对其他用户的影响与标签的不同主题及应用场景带来的负面影响等,鉴于此胡蓉[4]借助标签系统的优势结合主题模型,提出了一种基于标签-主题模型的标签推荐方法,邓晓懿等[5]结合用户评分和应用场景,提出了一种基于情境聚类和用户评级的协同过滤模型,王雪霞等[6]提出了基于评分和用户兴趣的协同过滤推荐

本文受安徽省教育厅自然科学重大项目(KJ2017ZD53),自然科学重点项目(KJ2017A799)资助。
万家山(1989—),男,硕士,讲师,CCF 会员,主要研究方向为数据分析和智能决策、模式识别、机器学习等。

方法，李瑞敏等[7]融合了标签和项目评分信息构建了用户-项目-标签的三维网络，在挖掘潜在语义的前提下采用二部图协同算法来实现推荐。

基于用户标签、评分或用户行为构建用户模型的研究较为普遍，但较少从用户画像的视角建立用户模型。目前用户画像构建方法普遍采用的是用户访谈式[8-10]。

用户画像也叫用户信息标签化，通过收集用户社会属性、偏好特征等维度数据，对用户或者产品特性进行刻画。用户画像是基于用户行为分析获得的对用户的一种认知表达，也是后续数据分析加工的起点。可作为定向广告投放和个性化推荐的前置条件，在学习资源推荐应用中，为了使推荐结果满足用户个性化的需要，往往根据用户的基础信息、学习资源信息、访问信息、行为偏好，以及隐式兴趣等维度建立用户模型，生成用户画像。用户画像通常是结合定性和定量分析展开描述的。

相对于用户定性画像，用户定量画像相对简单。在定量用户画像的建模过程中重点考虑用户画像的颗粒度，即用户画像应该如何细化。颗粒度越小，用户画像的描述越精确，进而影响用户模型的准确建立，有助于提高推荐系统的精确性。当然，用户画像也不是越小越好，颗粒度越小需要的用户个人数据就越多，一方面会提高用户建模的成本，同时也会导致用户画像的普适性降低。因此，用户画像的颗粒度需要依托于应用场景和用户需求的实时动态调整。

仅仅通过定量的方式是无法全面、精准地刻画用户画像的，如用户的兴趣模型很难用算术法则来表示一个人对学习资源内容的兴趣模型，以及用户会随着时间变化而发生怎样的转变。因此，需要结合定性和定量的方式将用户画像刻画出来。

2　用户建模与聚类

用户画像是用户建模的直观体现，不是简单的数字游戏，而是严肃的业务问题。构建用户画像的核心是利用标签进行建模[11]，每一个精炼的短标签背后是基于大量数据的分析挖掘，标签不是简单的短文本或符号，而是需要与具体业务紧密关联。

2.1　数据采集与处理

现今在线学习平台的种类众多，产品层次不齐，但很少有产品能够抓住核心之处：如何抓住用户的核心诉求，也就是抓住用户使用它的核心功能。以在线学习资源为例，核心需求为：寻找资源，即找到自己需要的资源。这需要系统能够智能化，能够对用户和学习资源有深度的理解。但一切智能都是建立在数据基础上的，需要有丰富的标签以及自然语言理解的能力，能够收集到用户的所有相关数据，包含静态信息数据和动态信息数据。

静态信息数据相对稳定，主要包括用户基本属性，这类信息一般为结构化信息，其收集方式主要为显式收集，最直接的方式就是通过表单要求用户提供相关个人信息。通过这种方式，推荐系统能够获取真实信息而无须过多地建模预测，将更多的精力放在数据清洗上。

动态信息数据是用户不断变化的行为信息，如观看视频资源、搜索信息、发表评分、互动交流等，这类信息一般为非结构化或半结构化的信息，收集方式主要为隐式收集。用户在推荐系统中的交互行为会产生很多相关数据，如用户搜索了某个资源、查看了资源的评论、收藏了某个主题、阅读浏览的时长等也会表明用户的感兴趣程度，系统会在不干扰用户与推荐系统交互的基础上，从用户的操作行为和上下文信息中获取相关数据，如图 1 所示。

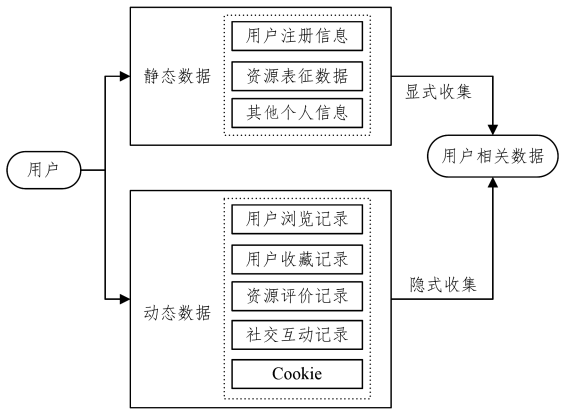

图 1　用户相关数据的收集

这种方式成为推荐系统中最主要的用户信息收集方式。一般地，隐式信息收集的准确性较显式信息收集方式差，但通过大量用户信息的收集和挖掘可相应提高准确性，或将隐式信息收集作为显式信息收集的一种补充，来缓解矩阵的稀疏性带来的冷启动问题的负面影响。

单纯的数据无法直接使用，必须通过对大量行为信息进行清洗、筛选来去除噪声数据，然后通过算法和模型学习建立动态的用户画像，实现用户共享才能带来巨大的数据应用价值[11]。

2.2　近邻算法 KD-Tree 实现

KNN 聚类算法（K-Nearest-Neighbors Classification），又叫 k 近邻算法。其核心思想是确定测试样本的类别，通过寻找所有训练样本中与该测试样本“距离”最近的前 k 个样本，然后判断这 k 个样本大部分的所属类别，换言之让最相似的 k 个样本来投票决定测试样本的类别。这里的距离，一般最常用的就是多维空间的欧氏距离。这里的维度指特征维度，即样本有几个特征就属于几维，一般 k 值的选择不会超过 20 个。KNN 聚类算法的关键环节有：1）相似度算法，常用的有欧氏距离、曼哈顿距离等；2）在计算相似度之前需要将数据归一化特征，比如使用 Max -Min 标准化，把所有特征都转换到 [0，1] 之间，转换公式为：

$$x' = (x - \min)/(\max - \min) \tag{1}$$

在 KNN 聚类算法中，计算对象间的距离并将其作为两两对象之间的非相似性指标，这里常见的距离计算方法一般使用欧氏距离或曼哈顿距离。

欧氏距离的计算式如下：

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2 \tag{2}$$

曼哈顿距离的计算式如下：

$$d(x,y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|} \tag{3}$$

其中，x_i 和 y_i 都是两个 n 维的特征向量。

从 KNN 聚类算法实现的过程可以发现，该算法存在两个严重的问题：1）存储结构不够优化，不利于后期的查询需要；2）计算量较大，计算非常耗时，因为需要计算每个待分类的样本数据与全体已知样本之间的距离，通过排序才能获得

该样本的 k 个最近点。

在 k 近邻算法的实现过程中，主要通过穷举搜索并计算距离，即计算当前未知样本与每一个已知样本的距离并进行快速 k 近邻搜索。计算并存储好以后，再查找 k 近邻。当训练样本量级较大时，计算非常耗时。为了提高 KNN 搜索的效率，考虑采用 KD-Tree 类型的数据结构存储样本数据，以减小计算距离的算法时间复杂度。

KD 树(K-dimension tree)是一种二叉树，主要针对 k 维数据空间进行层次划分，对 k 维空间中的样本点进行存储以便快速检索其树形数据结构[13-14]。利用 KD-Tree 可以省去对大部分样本数据点的搜索，从而减小搜索的算法时间复杂度。

KD-Tree 中的每个节点都是向量，这与一般的二叉树按照数值大小划分有所不同。KD-Tree 的每层需要选定向量中的某一维，然后根据这一维按左小右大的方式划分数据。“如何选择向量确定从哪一维度进行划分?”是过程中需要解决的关键问题。目前最为简单的解决方法是随机选择某一维或按顺序选择，但是更好的方法应该是根据方差来衡量分散的程度，在数据比较分散的那一维进行划分。KD-Tree 构建算法如算法 1 所示。

算法 1 构造平衡 KD-Tree

输入：k 维度空间数据集 Data_Set(以下简称 D)；所在的空间：Range

D=$\{x_1, x_2, \cdots, x_N\}$，其中 $x_i=(x_i^1, x_i^2, \cdots, x_i^k)$，i=1,2,…,N

输出：kd，类型为 KD-Tree

Step 1 If Data-set(D)为 NULL，则返回空 KD-Tree。

Step 2 确定 split(切割)域：对于描述的数据(特征矢量)，统计每个维度的数据方差。计算方差最大的维度，对应的为 split 域的值(数据方差大表明沿该坐标轴方向上的数据分散得比较开，在这个方向上进行数据分割有较好的分辨率)。

Step 3 确定根结点，生成 KD-Tree：

Begin：构造根结点。选择 x^1 作为坐标轴，为 D 中所有样本数据的 x^1 坐标的中位数作为比较对象(即切分点)，如果小于切分点的则划到 A 集合，大于则划入 B 集合。因此将根结点对应的超矩形区域切分为 A 和 B 两个集合。由根结点开始生成深度为 1 的二叉树：以坐标轴 x^1 进行划分，对于小于切分点的样本划分到左子结点(即 A 集合)，对于大于切分点的样本划分到右子结点(即 B 集合)，将正好落在切分平面上的样本点保存在根结点。

Repeat：构造平衡二叉树，对深度为 j 的结点，选择 x^{l_0} 作为切分的坐标轴，$l_0=(j \bmod k)+1$，以该结点所属集合中所有样本的 x^{l_0} 坐标的中位数为切分点，将该结点对应的超矩形区域切分为 A 和 B 两个子集合。由该结点生成深度为 j+1 的左、右子结点：左子结点对应坐标 x^{l_0} 小于切分点的 A 集合，右子结点对应坐标 x^{l_0} 大于切分点的 B 集合。

End：直到某个子集合不能再进行切分时，则将该集合中的数据保存到叶子结点。

2.3 用户画像的构建

用户画像主要包含定性画像和定量画像两个部分，其中定性画像主要包含用户的基本特征、行为特征、兴趣模型和学习资源表征等，定量画像主要包含用户的基础变量、兴趣偏好等可量化的数据特征。

用户画像的数据模型可以描述为：

标签权重＝衰减因子×行为权重×网址子权重

上述数据模型权重值的选取只是举例参考，具体实际操作过程中的权重值需要根据具体业务需求进行二次建模。因此，在用户画像确立过程中，先强调从整体思考构建初步的模型，然后再根据业务需求逐步细化模型。

用户画像的构建分为 3 个部分：1)需要收集用户的基础数据和用户行为数据作为构建用户画像需要的数据基础；2)需要将用户画像和具体业务结合，对用户画像的构建需要符合特定业务的需求；3)在综合考虑前两者的基础上进行数学建模，从已有的数据中进行深层次挖掘，结合具体的业务了解用户的需求，并通过数据可视化的方式将有用的信息展现出来，这才是用户画像的本质所在。

用户画像不同的应用领域有着差异的组成元素，如电商领域的用户画像主要侧重用户的消费习惯，而学习资源推荐领域的用户画像则主要侧重用户的学习兴趣和需要。

用户画像构建流程如图 2 所示。

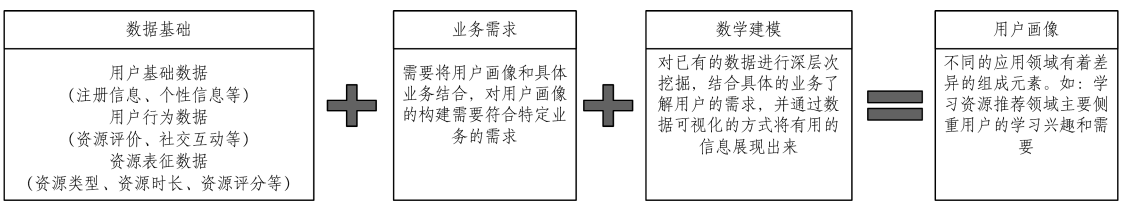

图 2 用户画像构建流程

3 在线资源推送框架

基于第 2 节用户画像的构建流程，进一步完善基于用户与资源画像的在线资源推送框架。

(1)用户偏好特征计算包括基于历史资源的用户偏好提取、偏好标签提取、其他个人信息(如专业、研究领域等)，结合用户行为特征数据构建用户画像。其中，偏好标签是通过历史相关用户数据进行采集和分析的，提取偏好标签信息，建立相对稳定的关键词向量，计算与标签较匹配的资源，然后将历史资源与标签融合，补充其他个人信息计算用户偏好。

(2)用户活跃性特征计算需要用户行为的活跃性度量、用户评论时间、用户行为互动性频率等指标，将用户转发资源、浏览资源、评论资源等行为的时间区间大小作为行为活跃性的指标。用户互动性通过用户评论的时间区间大小，计算用户间互动频率的紧密程度，共同确定用户互动性指标。

(3)资源表征的刻画需要基于资源结构和内容进行特征提取，融合资源的评论信息对资源进行定性和定量描述，构建资源画像。

在模型的训练过程中，通过用户偏好、用户活跃性、资源表征等信息进行计算，然后建立基于用户画像的在线资源推送框架，针对用户进行建模生成用户画像，从而进行相似度计算以确定目标用户，资源画像的构建是在资源结构、内容、评论等资源数据的基础上进行特征提取生成的。

在线资源推送框架如图 3 所示。

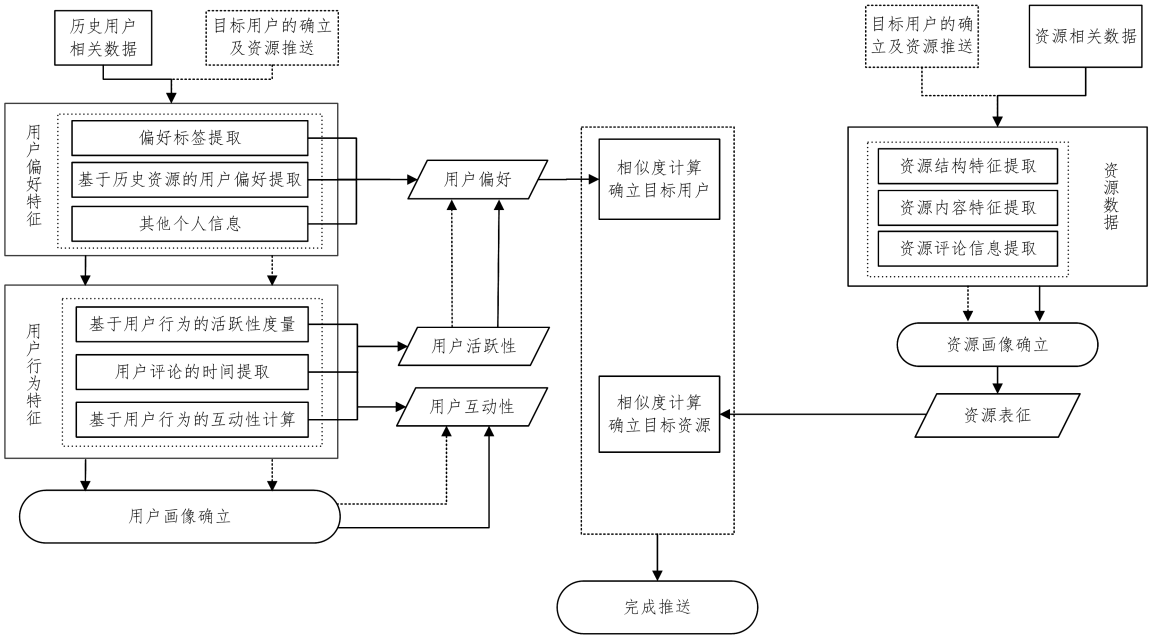

图 3　在线资源推送框架

4　实验与结果

本文借助博思智慧学习平台的学生数据来源，从手动分类的数据集中抽取了 1000 名学生用户数据集作为测试数据集。通过 KD-Tree、朴素贝叶斯（Naive Bayes）以及 BP 神经网络（BPNN）等算法，对比时间开销、聚类准确率和聚类召回率。

实验环境如下：Window 10 操作系统；CPU intel core i7-8550U 2.00 GHz；16 GB 内存。

实验 1　通过变化数据集大小（分别选择 200，500，800 和 1000 数据集的情况下）来比较聚类带来的时间开销。如图 4 所示，本文提出的 KD-Tree 聚类算法具有很好的性能优势，其考虑采用 KD-Tree 类型的数据结构存储样本数据，大大提高了 KNN 搜索的效率，进而减小了计算距离的算法时间复杂度。

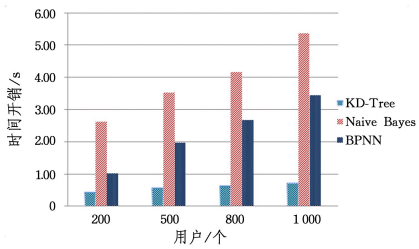

图 4　3 种算法的性能对比

实验 2　为了更好地展示算法的聚类准确率和聚类召回率，继续选择实验 1 中的 1000 名学生用户数据集作为测试数据集，具体聚类效果如表 1 所列。

表 1　3 种算法的聚类效果对比

算法	聚类	平均值
kd-tree	Precision	0.942
	Recall	0.942
Naive Bayes	Precision	0.875
	Recall	0.911
BPNN	Precision	0.893
	Recall	0.918

由表 1 可知，3 种算法的聚类准确率与召回率均较高，但是 KD-Tree 聚类算法的实验效果好于其他两种算法。由于这 3 种算法的聚类实验结果的差距较小，因此聚类算法的选取对最终聚类准确率和召回率的影响较小。

结束语　在大数据时代，系统要从大量的数据中了解用户，针对所有应用领域构建统一的推荐模型几乎是不可能的。本文借助在线学习平台资源推送的背景，研究如何使得在线学习平台更加智能、建立科学的用户画像和提出智能的资源推送算法是解决问题的关键。本文主要通过 KD-Tree 实现的 KNN 聚类算法分析用户偏好和行为特征，使得用户聚类变得更加快速。结合具体业务需求进行二次建模使得模型细化，聚类效果将更加精准。当前的资源推送框架还存在一些不足，在用户画像生成过程中，不仅仅包括用户与项目之间的交互数据和业务场景的影响，还应涉及到社会化关系影响、用户偏好的动态演化和资源特征的动态变化等，建模过程中考虑更多的要素往往能够提升推荐系统性能，希望本文的研究思路能为相关领域的学者和工程技术人员提供帮助。

参 考 文 献

[1]　张哲. 基于微博数据的用户画像系统的设计与实现[D]. 武汉：华中科技大学，2015.

[2]　王卫平，王金辉. 基于 Tag 和协同过滤的混合推荐方法[J]. 计算机工程，2011，37(14)：34-35.

[3]　刘凯鹏，方滨兴. 一种基于社会性标注的网页排序算法[J]. 计算机学报，2010，33(6)：1014-1023.

[4]　胡蓉. 基于标签—主题模型的标签推荐研究[D]. 武汉：华中师范大学，2013.

[5]　邓晓懿，金淳，韩庆平，等. 基于情境聚类和用户评级的协同过滤推荐模型[J]. 系统工程理论与实践，2013，33(11)：2945-2953.

[6]　王雪霞，李青，李季红. 基于共同评分项目数和用户兴趣的协同过滤推荐方法[J]. 计算机应用，2014，34(11)：3140-3143.

[7]　李瑞敏，林鸿飞，闫俊. 基于用户-标签-项目语义挖掘的个性化音乐推荐[J]. 计算机研究与发展，2014，51(10)：2270-2276.

（下转第 467 页）