文章编号:1000-2367(2020)06-0019-06 DOI:10.16366/j.cnki.1000-2367.2020.06.003

面向学生画像的偏好知识获取研究

王晓东,江培超,李梦莹,郝明丽,胡富珍

(河南师范大学 计算机与信息工程学院;教学资源与教育质量评估大数据河南省工程实验室,河南 新乡 453007)

摘 要:针对信息过载导致学生不能有效获取偏好知识的问题,提出一种面向学生画像的偏好知识获取方法.利用学生浏览知识内容,通过学生关键词、主题分布两个维度,构建学生画像向量空间模型.据此,计算学生与知识之间的相似度,获取直接偏好知识.利用学生浏览知识内容进行聚类分析,根据学生学习行为设计算法,获取间接偏好知识.以实际运行系统中提取的学生学习行为信息为实验数据,进行实验分析,结果表明,获取的偏好知识能更好地刻画学生画像.

关键词:学生画像;偏好知识;学习行为

中图分类号:TP399 **文献标志码**:A

大数据时代,教育领域对个性化学习的需求越来越高.教育大数据在走向多元化的过程中存在许多问题,如信息过载,这导致学生难以在海量数据中有效获取其偏好知识.如何准确刻画学生并充分描述学生的多维、多态特征,是解决此问题的关键.用户画像是真实用户的虚拟代表,是建立在一系列属性数据之上的用户模型.作为描述用户特征的工具,通过其获取用户偏好,已经取得了一定成果.文献[1]基于 Folksonomy 构建用户兴趣画像,文献[2]基于用户控制行为(主动性、耐光性等)构建用户画像,获取用户对灯光的照明偏好.用户因素、项目特征对用户偏好存在非线性影响,文献[3]使用深度前馈网络,构建用户画像和项目特征,预测用户对项目的满意度.由于学生身处复杂的学习环境并具有独特的学习行为,使其比一般用户更具特殊性,因此目前对学生画像的相关研究较少.肖君等[4]从知识特征、行为特征和态度特征 3 个维度设计在线学习者画像模型.王晓东等[5]利用学生特征(学习水平、学习风格等)构建学生模型.文献[6]基于用户阅读过程中产生的注释痕迹构建学习者画像.文献[7]基于模糊树构建学习者模型与学习活动模型,向学习者推荐学习活动.目前,学生画像模型大多是直接对学生的学习行为进行信息提取,没有对其深入分析并挖掘学生个性信息,不能有效获取学生的偏好知识,刻画学生画像.以移动自主学堂系统[8-9]数据为基础,利用学生阅读的知识内容,通过学生关键词、主题分布两个维度构建学生画像向量空间模型,挖掘学生偏好知识信息,细化学生画像,有助于个性化学习.

1 学生画像向量空间模型构建

学生可能会对与其阅读内容相似的知识感兴趣,也有可能对具有相似主题的知识感兴趣[10].为此,利用学生浏览知识内容,通过学生关键词、主题分布两个维度构建学生画像向量空间模型,如图 1 所示.给定知识集合 $R=\{R_1,R_2,\cdots,R_N\}$,学生集合 $S=\{S_1,S_2,\cdots,S_U\}$,R_n 表示知识 n,S_u 表示学生 u.对于学生 u,将其形式化表示为 $S_u=\langle F_u;G_u\rangle$,其中 F_u 表示学生 u 的关键词向量,G_u 表示学生 u 的主题分布向量.

1.1 知识预处理

根据移动自主学堂系统数据[8-9],构建学生画像向量空间模型,挖掘学生偏好知识信息,需进行预处理.

收稿日期:2019-08-04;**修回日期**:2020-05-17.

基金项目:横向研究项目(5201119160001;5202069169001);河南师范大学研究生科研创新项目(YL201917).

作者简介:王晓东(1963-),男,河南永城人,河南师范大学教授,博士,研究方向为本体工程、教育大数据,E-mail:wxd@htu.cn.

通信作者:江培超,E-mail:hnujpc@gmail.com.

将非文本类型的知识进行语义标注,从而将学生的浏览内容以文本形式呈现.已知知识集合 $R=\{R_1,R_2,\cdots,R_N\}$,对于知识 n,构建知识向量空间模型 $R_n=\langle K_n;E_n\rangle$.其中 K_n 表示知识 n 的关键词向量,E_n 表示知识 n 的主题分布向量.

1.1.1 知识关键词提取

对于知识集合 R 中的每个知识,使用 jieba 分词工具进行分词处理,使用的停用词典包括常见的中英文虚词、标点、HTML 占位符等,共计 3 974 项.然后,通过 TF-IDF 算法[11]计算分词后词语的权重.最后,构建知识关键词向量 $K_n=\{\langle K_{n1},\omega_{n1}\rangle,\langle K_{n2},\omega_{n2}\rangle,\cdots\}$.其中 K_{nj},ω_{nj} 分别表示知识 n 的关键词 j 及其对应权重.选择 $\omega\geq 0.02$ 的关键词表示知识.

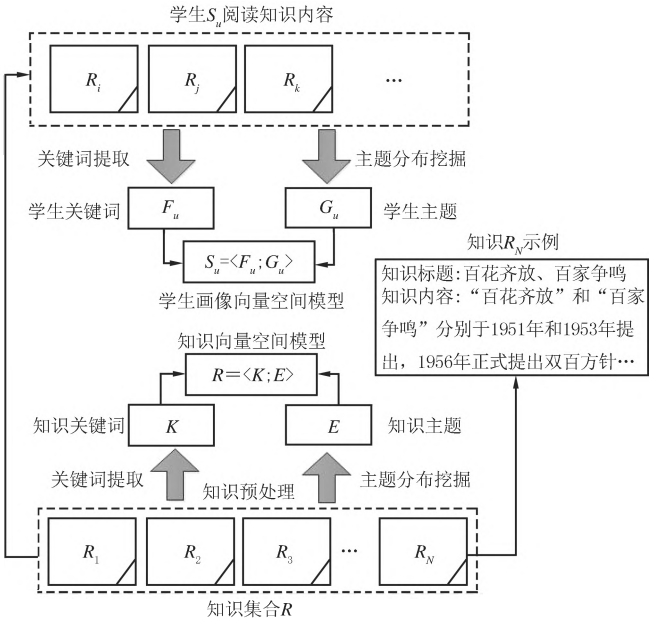

图1 学生画像向量空间模型构建

Fig. 1 Construction of vector model of student profile

使用 TF-IDF 算法计算权重 ω_{nj}.TF-IDF 分为词频(Term Frequency,TF)和逆文档频率(Inverse Document Frequency,IDF)两部分,两部分的乘积共同决定知识关键词的权重.词频计算公式为:$TF(n,j)=c(n,j)/s(n)$.其中,$c(n,j)$ 表示关键词 j 在知识 n 中的频数,$s(n)$ 表示知识 n 分词后的词语总数.逆文档频率计算公式为:$IDF(j)=\log(N/(l(j)+1))$.其中,N 表示知识集合 R 中知识的个数,$l(j)$ 表示 R 中包含关键词 j 的知识的数量.权重 ω_{nj} 计算为:$\omega_{nj}=TF(n,j)\times IDF(j)$.为使权重处于[0,1]区间内,使用余弦归一化的方式对权重进行处理:$\omega_{nj}=\omega_{nj}/\sqrt{\sum_{n=1}^{N}\omega_{nj}^2}$.

1.1.2 知识主题挖掘

对于 R 中的每个知识,使用 LDA 主题模型[12]挖掘潜在的知识主题,构建知识主题分布向量 $E_n=\{\langle E_{n1},\omega_{n1}\rangle,\langle E_{n2},\omega_{n2}\rangle,\cdots\}$,其中,$E_{nj},\omega_{nj}$ 分别表示知识 n 的主题 j 及其对应权重.

LDA 主题模型是一个具有“文本-主题-词”3 层结构的贝叶斯概率模型,它从知识语料库中提取代表性词语列表作为某一主题,最终将 R 中的每个知识的主题以概率分布的形式呈现.在 LDA 建模过程中,使用条件概率 $p(z_j\mid n)$ 表示知识 n 中的主题 z_j 的分布概率,使用条件概率 $p(w\mid z_j)$ 表示每个主题 z_j 中词语 w 的分布概率.词语 w 在知识 n 中的分布概率:$p(w\mid n)=\sum_{j=1}^{T}p(w\mid z_j)p(z_j\mid n)$.其中,$T$ 表示主题个数,选择 $T=20$ 保证知识主题表示的有效性.

1.2 学生关键词及主题提取

利用学生浏览的知识内容构建学生关键词向量.使用 jieba 分词工具对每个学生浏览的知识内容分词处理.通过 TF-IDF 算法计算分割结果中词语的权重,构建学生关键词向量 $F_u=\{\langle F_{u1},\omega_{u1}\rangle,\langle F_{u2},\omega_{u2}\rangle,\cdots\}$.其中 F_{uj},ω_{uj} 分别表示学生 u 的关键词 j 及其对应权重.选择 $\omega\geq 0.02$ 的关键词表示学生.

利用学生浏览的知识内容构建学生知识主题分布向量.使用 LDA 主题模型挖掘学生潜在知识主题,构建学生主题分布向量 $G_u=\{\langle G_{u1},\omega_{u1}\rangle,\langle G_{u2},\omega_{u2}\rangle,\cdots\}$.其中 G_{uj},ω_{uj} 分别表示学生 u 的主题 j 及其对应权重.选择 $T=20$ 保证学生主题表示的有效性.

2 偏好知识获取

根据构建的学生画像向量空间模型获取偏好知识,从而细化学生画像,如图 2 所示.(1)直接偏好知识获

取.将学生 u 的向量空间模型 S_u 与知识集合 R 中的每个知识进行相似度比对，选择与学生 u 相似度最高的前 top-k 个知识作为其直接偏好知识.(2)间接偏好知识获取.利用学生浏览内容对学生聚类分析，将阅读内容相似的学生聚为一类.根据聚类结果，分析学生学习行为设计算法，间接获取学生 u 的 k 个偏好知识.(3) 将步骤(1) 与(2) 获取学生 u 的 k 个直接和间接偏好知识共计 L 作为其偏好知识，其中 $L=2k$.细化学生画像 $S_u=\langle P_u\rangle$，$P_u=\{R_i \mid i\in 1,2,\cdots,N\}$ 且 $|P_u|=L$.其中 P_u 表示学生 u 的偏好知识，R_i 表示知识 i，N 表示 R 中知识的个数.细化后的学生画像反映了学生对知识的偏好.

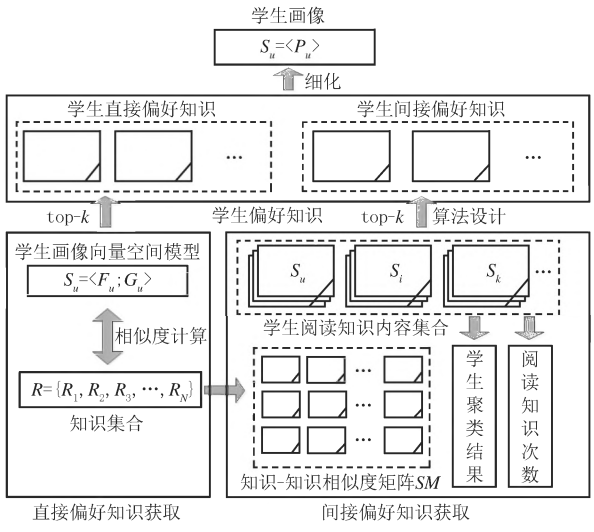

图2　偏好知识获取流程

Fig.2　The process of preference knowledge acquisition

2.1　直接偏好知识获取

使用 Jaccard 公式计算学生 u 与知识集合 R 中每个知识的相似度，与学生 u 相似度最高的 k 个知识作为学生 u 的直接偏好知识，相似度计算如(1)～(3)式所示.

$$S(S_u,R_n)=\frac{\alpha S(F_u,K_n)+(1-\alpha)S(G_u,E_n)}{\sqrt{\alpha^2+(1-\alpha)^2}},\tag{1}$$

$$S(F_u,K_n)=\frac{|I(F_u,K_n)|}{|U(F_u,K_n)|},\tag{2}$$

$$S(G_u,E_n)=\frac{|I(\max(G_u),\max(E_n))|}{|U(G_u,E_n)|},\tag{3}$$

其中，$S(S_u,R_n)$ 表示学生 u 与知识 n 的相似度，$I(F_u,K_n)$ 表示学生关键词 F_u 与知识关键词 K_n 的交集，$U(F_u,K_n)$ 表示学生关键词 F_u 与知识关键词 K_n 的并集.$\max(G_u)$ 表示从学生 u 的主题分布向量 G_u 中选择最高主题权重 ω 所对应的主题，主题中包含了一系列代表性词语，$\max(E_n)$ 同理.α 为权重参数，令其为 0.70，使得获取效果最好.

2.2　间接偏好知识获取

学生之间阅读的知识内容存在相似性，相似性高的一类学生可能具有相同偏好.为此，分析与目标学生阅读知识内容相似的学生，来反映目标学生的偏好知识.根据学生浏览的知识内容对学生聚类分析，将阅读内容相似的学生聚为一类.根据聚类结果，分析学生的学习行为设计算法，获取间接偏好知识.

2.2.1　学生聚类

将学生浏览的知识内容载入语料库，使用 jieba 分词工具对知识语料进行分词处理.构建词袋向量 ω_M 并使用 IF-IDF 算法计算每个词汇的权值，得到矩阵 $\boldsymbol{D}=\{\omega_{ij}\}_{U\times M}$，如(4)式所示：

$$\boldsymbol{D}=\begin{bmatrix}\omega_{11} & \cdots & \omega_{1M}\\ \vdots & & \vdots\\ \omega_{U1} & \cdots & \omega_{UM}\end{bmatrix},\tag{4}$$

其中，矩阵行数 U 表示学生数量，列数 M 表示分词后语料库中词汇的数量，每一行表示不同学生的词袋向量，ω_{ij} 表示第 i 个学生的词袋向量对应第 j 个词汇的权重.由于词汇过多可能导致 D 为稀疏矩阵，采用主成分分析法[13](Principal Component Analysis，PCA)对矩阵降维处理得到矩阵 $D'=\{\omega_{ij}\}_{U\times M'}$，其中 $M'<M$.最后，使用 k-means 算法对 D' 聚类.

已知学生集合 $S=(S_1,S_2,\cdots,S_U)$,其中 $S_i=\{\omega_{i1},\omega_{i2},\cdots,\omega_{iM}\}$ 表示学生 i 的词袋向量.将学生集合 S 分为 k 组 $C_1,C_2,\cdots,C_k$,具备的性质有 ①$C_i\neq\varnothing$,$i=1,2,\cdots,k$;②$C_i\cap C_j=\varnothing$ 且 $\bigcup_{i=1}^{k}C_i=C$,$i,j=1,2,\cdots,k$ 且 $i\neq j$.

k-means 算法主要有以下 4 个步骤:(1)令 $H=1$,从学生集合 S 中随机选取 k 个点($Q_{1(H)},Q_{2(H)},\cdots,Q_{k(H)}$)作为 k 个簇的中心;(2) 当且仅当满足 $\|S_i-Q_j\|<\|S_i-Q_t\|$($j=1,2,\cdots,k$ 且 $j\neq t$),则将 $S_i(i=1,2,\cdots,U)$ 归入簇 $C_j(j=1,2,\cdots,k$ 且 $j\neq t)$;(3) 计算簇的新中心点($Q_{1(H+1)},Q_{2(H+1)},\cdots,Q_{k(H+1)}$),计算公式为:$Q_{k(H+1)}=\dfrac{1}{m_k}\sum_{S_j^k\in C_k}S_j^{(k)}$.其中,$m_k$ 是处于簇 C_k 的点的数量,且令平均误差函数:$F(H+1)=\sum|S_j^{(k)}-Q_{k(H)}|^2/\dfrac{1}{m_k}$;(4) 给定算法精度 δ,如果 $|F(H+1)-F(H)|<\delta$ 则算法结束,否则 $H=H+1$,返回步骤(2)继续.

2.2.2 偏好知识获取算法

阅读知识内容相似的学生可能具有相同的偏好,若学生对某一知识阅读频繁,那么该知识可能是学生的偏好知识.为此,构建知识-知识相似度矩阵(Similarity Matrix,***SM***),在学生聚类结果的基础上,根据 ***SM*** 以及阅读知识内容、阅读知识次数等学习行为设计算法,间接获取学生偏好知识.***SM*** 如(5)式所示,偏好知识获取算法见表 1.

表 1 偏好知识获取算法

Tab.1 The algorithm of preference knowledge acquisition

INPUT:(1)学生聚类结果 $C=\{C_1,C_2,\cdots,C_f\}$;(2)C_k 中学生集合 $C_k=\{S_1,S_2,\cdots,S_j\}$,$S_j=\{\langle R_{j1},C_j^{R_{j1}}\rangle,\langle R_{j2},C_j^{R_{j2}}\rangle,\cdots,\langle R_{jn},C_j^{R_{jn}}\rangle\}$,其中 S_j 表示学生 j 阅读知识的集合,R_{jn} 表示学生 j 阅读的知识 n,$C_j^{R_{jn}}$ 为其对应的阅读次数,$C_k\in C$;(3)***SM*** 矩阵.

OUTPUT:学生 j 的间接偏好知识

1:For C_k in C do
2: For S_j in C_k do
3: For $i=1$:end do //对于 C_k 中除学生 j 以外的其他学生
4: In[i][j]=$S_i\cap S_j$($S_i\in C_k$,$i\neq j$)//学生 i 与学生 j 阅读知识的交集
5: Totp10In[i][j]=Sort(In[i][j])//选择学生 j 在 In[i][j]中阅读次数最多的 10 个知识
6: TotalCount[i][j]=ADD($C_j^{\text{In}[i][j]}$)//学生 j 阅读 In[i][j]中知识的总次数
7: AvgCount[i][j]=$\dfrac{\text{TotalCount}[i][j]}{\text{Length}(\text{In}[i][j])}$//学生 j 阅读 In[i][j]中知识的平均次数
8: For each R_{in}($R_{in}\in S_i$,$R_{in}\notin$ In[i][j]) do
9: $\text{AvgSim}[R_{in}]=\dfrac{\text{ADD}(S(R_{in},R_{\text{Top10In}[i][j]}))}{10}$// 知识 R_{in} 与 Top10In 中知识的相似度相加之和除以 10
10: $\text{Weigh}[R_{in}][j]=\dfrac{\text{AvgSim}[R_{in}]}{\text{AvgCount}[i][j]}$// 学生 j 对知识 R_{in} 的偏好权重
11: End
12: End
13: Sort(Weigh)//选择权重最大 top-k 的个知识为学生 j 的间接偏好知识
14: End
15:End

$$\boldsymbol{SM}=\begin{bmatrix}S(R_1,R_1) & \cdots & S(R_1,R_N)\\ \vdots & & \vdots\\ S(R_N,R_1) & \cdots & S(R_N,R_N)\end{bmatrix},\tag{5}$$

由(5)式可知,***SM*** 为对称矩阵,矩阵的值为知识之间的相似度.利用 1.1.1 节构建的知识关键词向量,使用

Jaccard 公式计算知识 R_u 与 R_f 的相似度：$S(R_u,R_f)=\dfrac{|I(K_u,K_f)|}{|U(K_u,K_f)|}$.

3　实验分析

3.1　实验数据与标准

从移动自主学堂系统[8−9]中提取 200 个学生的学习行为信息及相应的 4 119 个知识作为实验数据，构建学生画像向量空间模型，获取学生偏好知识.数据中包括学生浏览知识内容、知识信息、学生收藏知识等.硬件 MAC OS 平台：英特尔酷睿 i5，8 GB 内存，1 TB 硬盘，使用 python3 进行实验.选用 $F1$ 值作为验证标准，对查准率与查全率进行整体评价，$F1$ 值计算如(6)式所示：

$$P=\frac{N}{L},R=\frac{N}{F},F1=\frac{2\times P\times R}{P+R},\tag{6}$$

其中，P 为查准率，R 为查全率，N 为准确获取学生偏好知识的个数，L 为实际获取学生偏好知识的个数，F 为学生收藏知识个数.

3.2　实验结果分析

面向学生画像的偏好知识获取受学生聚类个数 k 的影响，不同的聚类个数导致实验结果不同.同时，偏好知识获取个数 L 对 $F1$ 值有较大影响，取值太小则无法说明方法的有效性，取值过大会造成结果难以预估.抽取 50 名学生，在 L 一定的情况下，观察不同聚类个数 k 对 $F1$ 值的影响，如图 3 所示.

由图 3 可知，当 $2\leq k\leq 8$ 时，不同偏好知识获取个数 L 的 $F1$ 值较低且变化不稳定，实验效果并不理想，这可能是因为较少的聚类个数无法有效区分不同学生的偏好知识，导致其查准率、查全率较低，从而影响 $F1$ 值.当 $8<k\leq 16$ 时，不同 L 的 $F1$ 值呈上升趋势.$k=16$，$L=40$ 时，$F1$ 值最高，为 0.61.当 $k>16$ 时，不同 L 的 $F1$ 值呈下降趋势，这可能是因为随着聚类个数的增加导致学生区分过度，在学生聚类簇中，一些可以反映目标学生偏好知识的学生被划分到其他聚类簇中，导致查准率、查全率降低，从而 $F1$ 值降低.

最后，设定学生聚类个数 $k=16$，分析比较基于用户的协同过滤、基于内容、面向学生画像 3 种偏好知识获取方法在不同 L 下的 $F1$ 值，如图 4 所示.

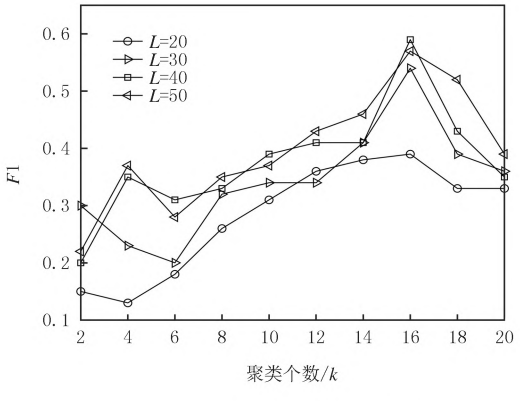

图3　不同聚类个数对F1值的影响

Fig.3　The influence of different cluster number on F1 value

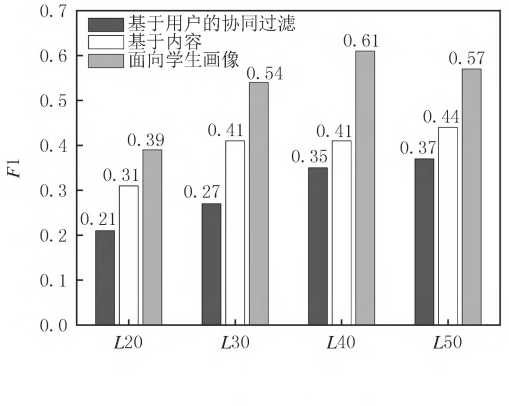

图4　偏好知识获取结果

Fig.4　The results of preference knowledge acquisition

由图 4 可知，面向学生画像的偏好知识获取效果最好.当 $L=40$ 时 $F1$ 值最高，为 0.61，分别比基于用户的协同过滤、基于内容的方法提高了 0.26、0.20.因此，可以证明获取的偏好知识能更好地刻画学生画像.

4　结束语

论文提出一种面向学生画像的偏好知识获取方法.通过学生关键词、主题分布两个维度构建学生画像向量空间模型，获取直接偏好知识.在学生聚类的基础上，通过学生学习行为设计算法，获取间接偏好知识.将直接偏好知识与间接偏好知识作为学生偏好知识，细化学生画像.通过实验分析，验证获取的偏好知识能更

好地刻画学生画像.后续工作将学生的认知能力特征融入学生画像模型.

参 考 文 献

[1] GOEL S,KUMAR R.Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels[J].Neurocomputing,2018,315(13):425-438.

[2] DESPENIC M,CHRAIBI S,LASHINA T.Lighting preference profiles of users in an open office environment[J].Building and Environment,2017,116(3):89-107.

[3] PURKAYSTHA B,DATTA T,ISLAM M S.Rating prediction for recommendation:constructing user profiles and item characteristics using backpropagation[J].Applied Soft Computing,2019,75:310-322.

[4] 肖君,乔惠,李雪娇.基于 xAPI 的在线学习者画像的构建与实证研究[J].中国电化教育,2019,384(1):128-134.
XIAO J,QIAO H,LI X J.The construction and empirical research of the learners' persona based on xAPI[J].China Educational Technology,2019,384(1):128-134.

[5] 王晓东,时俊雅,李淳.学习资源精准推荐模型及应用研究[J].河南师范大学学报(自然科学版),2019,47(1):26-32.
WANG X D,SHI J Y,LI C.Accurate recommendation model and application of learning resources[J].Journal of Henan Normal University(Natural Science Edition),2019,47(1):26-32.

[6] OMHENI N,KALBOUSSI A,MAZHOUD O.Computing of learner's personality traits based on digital annotations[J].International Journal of Artificial Intelligence in Education,2017,27(2):241-267.

[7] WU D,LU J,ZHANG G.A fuzzy tree matching-based personalized e-learning recommender system[J].IEEE Transactions on Fuzzy Systems,2015,23(6):2412-2426.

[8] 王瑞,李永波,王晓东.移动自主学堂及其应用[J].河南师范大学学报(自然科学版),2014,42(6):162-166.
WANG R,LI Y B,WANG X D.Mobile autonomous school and its application[J].Journal of Henan Normal University(Natural Science Edition),2014,42(6):162-166.

[9] 王瑞.信息化环境下移动课堂教学模式探究[J].中国教育学刊,2015,12:59-62.
WANG R.Research on mobile classroom teaching mode in information environment[J].Journal of the Chinese Society of Education,2015,12:59-62.

[10] ZHU Z,LI D,LIANG J.A dynamic personalized news recommendation system based on BAP user profiling method[J].IEEE Access,2018,6:41068-41078.

[11] SALTON G,BUCKLEY C.Term-weighting approaches in automatic text retrieval[J].Information Processing & Management,1988,24(5):513-523.

[12] BLEI D M,NG A Y,JORDAN M I.Latent dirichlet allocation[J].Journal of Machine Learning Research,2003,3:993-1022.

[13] ABDI H,WILLIAMS L J.Principal component analysis[J].Wiley Interdisciplinary Reviews:Computational Statistics,2010,2(4):433-459.

Preference knowledge acquisition for student profile

Wang Xiaodong,Jiang Peichao,Li Mengying,Hao Mingli,Hu Fuzhen

(College of Computer and Information Engineering;Big Data Engineering Lab of Teaching Resources & Assessment of Education Quality,Henan Province,Henan Normal University,Xinxiang 453007,China)

Abstract:In order to solve the problem that information overload causes students not to effectively acquire their preference knowledge, this paper proposes a method to obtain preference knowledge for the student profile. On the basis of the students' browsing content, the vector space model of student profile is established through the following two dimensions: keyword and topic distribution. Based on this, the similarity between students and knowledge is calculated to obtain direct preference knowledge. Subsequently, students' browsing knowledge content is applied to conduct cluster analysis, and the algorithm is designed according to students' learning behaviors to obtain indirect preference knowledge. The learning behavior information extracted from the actual operating system is taken as experimental data. Experimental results reveal that the acquired preference knowledge can better depict the student profile.

Keywords:student profile;preference knowledge;learning behavior

[责任编校 陈留院 赵晓华]