计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):212−216 [doi: 10.15888/j.cnki.csa.005935]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

远程培训中个性化学习资源推荐算法①

冯亚丽, 于瑞芳, 孙龙安, 宋新起

(东北石油大学 计算机与信息技术学院, 大庆 163000)

摘　要: 本文提出一种基于标签的多因素推荐算法. 用户可以根据自己的需求, 进行因素自定义和优先级排序, 算法先根据用户初始化信息选取资源, 随后分析用户行为数据更新用户所属的群及用户的喜好, 再通过用户与项目相似度计算、项目关联度计算为用户推荐所需资源. 算法模型采用分类组合得出结果, 降低了相似度计算的复杂度. 将算法应用于企业远程培训平台的个性化学习模式中, 结果表明, 该算法较好地改善了用户个性化学习资源的推荐效果.

关键词: 推荐算法; 多因素; 教育模式; 标签; 远程教育

引用格式: 冯亚丽,于瑞芳,孙龙安,宋新起.远程培训中个性化学习资源推荐算法.计算机系统应用,2017,26(8):212–216. http://www.c-s-a.org.cn/1003-3254/5935.html

Recommendation Algorithm of Personalized Learning Resources in Remote Training

FENG Ya-Li, YU Rui-Fang, SUN Long-An, SONG Xin-Qi

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163000, China)

Abstract: This paper puts forward a recommendation algorithm of various factors based on the labels. Users can define the factors and sort them with priority according to their own requirements., This algorithm will select the resources according to the initial information of users, then update users' groups and users' preferences, then recommend the useful resources for users based on the similarity calculation between users and project and the correlation calculation of the project. The algorithm model adopts classification of combination to get the results, and reduces the complexity of similarity calculation. The algorithm was applied to the personalized learning platform of remote training platform in enterprise. The results show that this algorithm has much improved the recommendation effects of user`s personalized learning resources.

Key words: recommendation algorithm; various factors; education mode; label; distance learning

1　引言

在当今的信息化时代, 远程培训作为一种便捷、高效的培训模式迅速普及, 成为各企事业部门的数字化建设中的必选项. 远程培训平台在为用户提供越来越多的学习资源的同时, 也使用户在海量学习资源中获取所需的学习资源十分困难. 如何满足用户的个性需求, 成为远程培训平台建设中必须关注的问题, 也是越来越多学者研究的热点.

目前推荐算法的应用是解决用户个性需求的主要途径, 主要有协同过滤推荐、基于内容的推荐和基于规则的推荐等算法.

协同过滤(Collaborative Filtering Recommendations, 简称 CFR)推荐能过滤难用机器自动过滤的、基于内容分析的信息, 能对基于一些复杂的、难以表达的概念(信息质量、品质)进行过滤, 能挖掘用户的潜在兴趣, 推荐新颖的学习资源[1], 但存在用户对学习资

① 基金项目: 黑龙江省教育科学规划课题(十二五)(GBC1213052); A7 中国石油天然气集团公司工程技术生产运行管理系统; 省教育厅教学改革项目(JG2014010640); 云计算理念打造大庆教育云的研究(DSGB2016053)
收稿时间: 2016-12-11; 采用时间: 2017-01-12

源评价的稀疏性、新用户没有评价过的资源、新资源的最初评价和冷开始等问题[2].

基于内容推荐能为有特殊兴趣爱好的用户推荐新的或并不是很流行的项目, 无稀疏和冷开始问题; 但要求学习资源的内容能容易地被抽取成有意义的、有良好结构性的特征内容, 要求用户的需求必须用内容特征形式表示, 不能显式地得到其它用户的判断情况, 有信息挖掘不充分、只能处理文本型文件和缺少用户反馈等缺点[2].

基于关联规则的推荐可挖掘不同学习资源间的相关性, 为用户推荐相关学习资源. 缺点一是关联规则的发现既关键又耗时, 二是资源名称的同义性问题. 这些问题随着资源不断的增长, 使规则的制定和维护越加困难[刘志勇][2].

以上广泛应用三种推荐算法有优点, 但都有不足, 如协同过滤推荐的冷开始问题, 基于内容推荐的内容抽取问题, 基于关联规则推荐的关联规则和资源同义性问题, 在实际应用中达不到理想的效果. 针对企业远程培训的特殊性, 从解决推荐的冷开始问题着手, 兼顾抽取内容、发现关联规则和解决资源同义性提出一种基于标签的多因素推荐算法, 该推荐算法有效的弥补了主流推荐算法的不足, 并有效的降低了算法的时间复杂度.

本文的总体研究思路是首先从两个方向同时入手, 用户信息和资源信息, 通过对用户和资源的静态信息识别, 对用户和资源分类, 添加初始标签, 记录相似度、关联度. 然后根据动态行为信息, 添加或修正标签, 修改相似度、关联度. 通过标签的使用解决主流推荐的冷开始、内容抽取、资源同义性及规则发现问题.

2 基于推荐算法的学习者喜好及需求确定

2.1 基于用户的 CFR 确定学习者所属的群

远程培训平台面向的学习者众多, 要快速地向学习者推荐其所需的资源, 其中一个重要的问题是将学习者分群, 即具有相同喜好的学习者为一个群. 可根据用户在学习平台上曾经的上网行为或本次上网的行为, 利用协同过滤推荐算法具有挖掘用户的潜在兴趣, 推荐新颖的学习资源的特点来确定学习者所属的群.

协同过滤算法主要有基于用户和基于资源的 CFR.

基于用户的 CFR 的原理是分析用户喜好在用户群中找到与指定用户相似(或相同喜好)的用户, 综合统计分析相似用户对共同关注的某一信息评价. 估算出该指定用户对此信息的喜好程度. 依据对各信息的喜好程度阀值等因素, 确定相似喜好的用户子集(用户子群). 其形式化的描述为, 存在用户集合 User=(u_1, u_2, …, u_n)(u_i, 任一用户, i=1, 2, …n), n 个资源集合 Res=(r_1, r_2, …, r_n)(r_i, 任一资源, i=1, 2, …n), 就一定存在着对某些资源的喜好度大于给定阀值的相似用户子集 $\text{User}_{\text{sim}} \in$ User. 用函数 Sim(u_i, u_j)(i=1, 2, …n, j=1, 2, …n, $i \neq j$)描述 u_i 和 u_j 的相似度. 当 u_i 对资源 r_k 感兴趣时, u_j 也可能对 r_k 感兴趣. 在协同过滤算法领域对相似度的计算, 用户间相似度 Sim(u_i, u_j)计算依据为评分矩阵, 评分矩阵为 m 个用户对 n 个资源的访问形成的 m*n 矩阵, 它反映了用户的喜好, 其中 $R_{i,k}$ 标识用户 i 对资源 k 的评分.

目前, 在协同过滤算法中常用的相似度计算方法有皮尔逊相关数法(COR)、向量余弦法(COS)、调整的向量余弦法(ACOS)、约束的皮尔逊相关系数法(CPC)、斯皮尔曼相关系数法(SRC)等. 如果所研究的企事业单位的学习平台相对于其评分矩阵中的稀疏(sparsity)因子较少, 矩阵元素量较少, 可采用皮尔逊相关系数法(COR), 见公式(1):

$$sim(i,j)=\frac{\sum_{k\in I_{ij}}(R_{i,k}-\bar{R}_i)(R_{j,k}-\bar{R}_j)}{\sqrt{\sum_{k\in I_{ij}}(R_{i,k}-\bar{R}_i)^2}\sqrt{\sum_{k\in I_{ij}}(R_{j,k}-\bar{R}_j)^2}} \quad (1)$$

2.2 推荐学习者所喜欢的资源

2.2.1 基于资源的 CFR

基于资源的 CFR 是用函数 Sim(r_i, r_j)表示资源集合 Res=(r_1, r_2, …, r_n)中 r_i 和 r_j 的相似度. 对于用户来说, 当 r_i 和 r_j 的相似度大于阀值时, 访问了资源 r_i 的用户也会对资源 r_j 感兴趣. 因此, 可根据资源的相似度为用户推荐其可能喜欢的其它资源.

2.2.2 基于内容的推荐

基于内容的推荐(Content-based Recommendations 简称: CB), 是依据用户过去喜欢的资源为用户推荐和其曾经喜欢资源的相似资源, 计算方法是通过计算用户偏好和项目(Item, 表示资源)描述之间的相似性, 为用户推荐候选资源. 对于用户偏好和项目相似性的计算. 目前流行的是 TF-IDF(Term Frequency-Inverse Document Frequency)方法. 它是一种统计方法, 用以评估一个字词对于一个文档集, 或一个语料库中某份文件的重要程度. 字词的重要性与它在文档中出现的频率成正比, 同时又与它在语料库中出现的频率成反比. 程度高的词具有很好的类别区分能力, 适合于分类. 依

此算法, 可将资源分类, 有的资源可属多个类.

由于小型企事业的培训资料受到职域及专业的限制, 其数据通常被限定在企事业的职能域内. 因此, 可采用基于用户的 CFR 确定用户所属的群, 基于资源的 CFR 为用户推荐类似资源, 基于内容 CB 将资源分类; 一个用户可以属于不同的群, 一个资源也可以属于多个资源类, 一个用户群可以被推荐多个资源类; 用户及资源的推荐关系模型如图 1 所示.

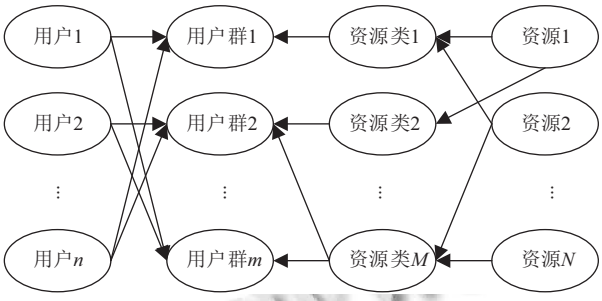

图 1 用户及资源的推荐关系模型

图 1 模型的方式避免了混合推荐带来的复杂计算问题. 但该方法受到用户数量及学习资源量限制的同时, 还很难实现用户的个性化学习需求. 为此本文在图 1 模型的基础上提出了基于标签的多因素推荐算法.

3 基于标签的多因素推荐算法

基于标签的多因素推荐算法是指用户可根据自身需求, 进行因素自定义和优先级排序. 在基于标签的多因素算法中, 标签标识 Item(资源)的目标、分类及内容; 类别标签与属性(因素)间相互依赖, 类别标签是资源某些属性子集的特征.

Item 用属性标注的方法描述, 属性标注使标签便于选择. 每个 Item 可拥有多个不同的标签, 这使 Item 拥有标识的同时具有语义信息. 推荐算法采用协同过滤内容推荐算法中的基本思想, 用户聚类信息由采用用户评分矩阵改为用户偏好信息进行聚类, 用户偏好信息与 Item 信息的相似性采用属性语义相似的计算方法.

针对培训平台上学习者的特点及资源特点, 本文构建了基于标签的多因素推荐模型, 如图 2 所示. 该模型主要分为 User 信息、Item 信息描述, 原始数据分析, Item 推荐等功能模块. 如为 User 推荐 Item 信息的过程是从数据层调用信息进入数据分析层进行 User 行为分析、User 间相似度计算、User 与 Item 相似度计算、Item 间关联度计算、User 偏好信息分析通过以上分析计算得到 Item 集合, 在通过 Item 的评分、过滤规则将符合条件的 Item 推荐给 User. 也就是 通过各个模块的协同工作, 为 User 推荐 User 喜欢的 Item.

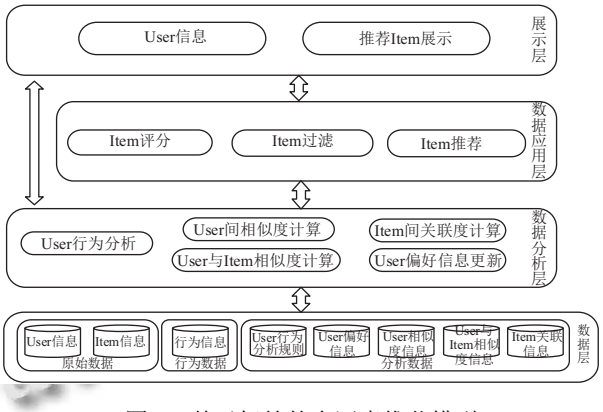

图 2 基于标签的多因素推荐模型

3.1 多因素描述

在该模型中, 数据主要分为原始数据、行为数据和分析数据 3 种资源信息. 原始数据为主要包括 Item 属性信息、User 信息等不需要加工的信息; 行为数据主要是 User 操作产生的行为记录信息; 分析数据包括 User 行为分析规则、User 偏好、User 相似度、User 与 Item 相似度、Item 关联度等信息.

3.1.1 原始数据

原始数据是 User 和 Item 在创建时赋予给 User 及 Item 的描述信息, 是实体本来就具有的. User 的原始属性, 如 User 所属公司类型、角色、岗位等, 如图 3 所示. Item 原始属性, 如: Item 类型(视频、文档等)、用途、该 Item 指导了工作流程中的什么阶段、Item 的内容是什么、适合哪个角色或者岗位来学习、时长或字数等, 如图 4 所示. 此类原始数据不需要加工. User 和 Item 的原始属性都采用标签进行标注, 标签的粒度更小, 可以让匹配更加精准.

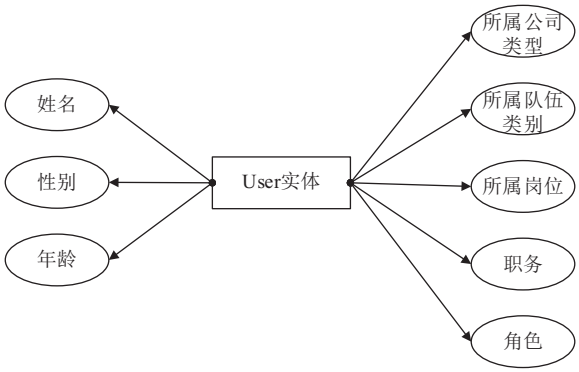

图 3 User 实体

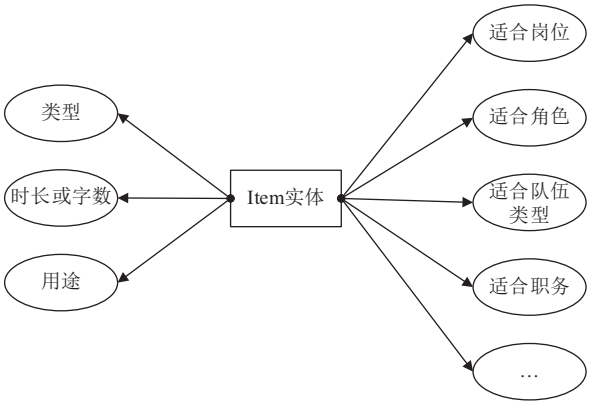

图 4 Item 实体

3.1.2 行为数据

行为数据是 User 在使用远程运维服务平台时所产生的操作数据, 如记录 User 登录时间, 以及 User 操作行为产生的数据, 如浏览了哪个教程, 该教程是否浏览完成, 搜索的关键词, 在线频率及时长等数据. 用户在浏览完 Item 后, 会给予 Item 进行 1-5 星的评分, 以表示 Item 的质量.

3.1.3 用户偏好信息

User 偏好信息可以通过初始配置和模式学习两种方法结合. 即 User 初始配置后, 随着 User 访问记录而动态更新的方式进行 User 偏好信息的管理. 为了更好评估 User 偏好信息, 可以对 Item 用属性, 属性值、权重三元组进行表示. 如 profile={<p_1, v_1, w_1>, <p_1, v_2, w_2>, …<p_m, v_n, w_x>}, 其中, 属性值被赋予特定的标签, 偏好信息随着 User 访问记录频次的增多而进行更新, 具体更新方法及值由行为数据分析得来.

3.1.4 用户相似度信息

User 相似度信息用于说明 User 之间的相似性, 系统中采用<User, 相似 User, 相似度>三元组进行表示. 可以根据 User 的属性中设置的标签, 进行 User 相似度判断, 标签重复越高, 相似度越高, 采用定时更新 User 相似度.

3.1.5 用户对项目偏好度信息

Use 对某个 Item 偏好度采用<User, Item, 相似度>三元组进行表示. 通过 User 属性的标签和 Item 属性的标签进行对比, 重复越高偏好度也越高, 采用定时更新.

3.1.6 项目关联匹配

Item 之间的关联度由属性标签进行比对得出. 采用<Item, 属性, 相似度>三元组进行表示. 标签重复越高, 表示关联越紧密.

3.2 原始数据分析

原始数据分析的主要功能是在实体描述的基础数据上进行分析处理, 主要通过 User 的历史访问信息进行偏好度更新、User 间相似度计算、User 与 Item 相似度计算等功能.

3.2.1 偏好度更新

偏好度更新功能主要是通过分析 User 对 Item 的访问历史来更新用户的偏好配置信息, 用户每次浏览 Item 都会被记录, 浏览到什么程度也会被记录, 并被赋予不同的权值. 偏好信息将按照下述步骤进行更新:

① 读取 User 信息;

② 读取 User 浏览 Item 的历史信息;

③ 判断用户操作获取喜好信息;

④ 读取偏好度分析规则;

⑤ 通过 Item 和 User 实体获取分析属性及属性值;

⑥ 通过属性值中的多种标签和偏好分析规则分析出偏好信息对;

⑦ 获取 User 已有偏好信息;

⑧ 保留初始偏好属性信息的前提下, 更新 User 偏好属性中的标签即 User 偏好项.

3.2.2 用户间相似度计算

User 相似度的计算是根据 User 的偏好信息进行用户聚类, 该相似度是基于 User 协同过滤的基础, 采用 User 偏好信息为聚类数据, 使系统可以根据该数据进行协同推荐. 在根据相同偏好的 User 聚类中, 将偏好信息看作项目信息, 以基于 User 的余弦相似性进行 User 的聚类, 在相同 User 访问一个项目时, 同时向同组 User 推荐该 User 的访问项. User 间相似度计算的主要过程如下:

① 读取所有 User;

② 根据 User 进行循环;

③ 获取每个 User 的偏好属性及其标签值, 并根据标签值进行偏好度比对;

④ 如果大于阀值;

⑤ 增加 User 间相似度.

3.2.3 用户与项目相似度

由于 User 配置信息采用了具有语义信息的属性作为偏好配置, 在 User 和 Item 相似度计算时, 采用基于属性相似性的计算方法, 该计算过程如下:

① 读取所有 User;

② 读取所有 Item;

③ 循环获取 User 的偏好属性及标签值;

④ 循环获取 Item 相关属性及标签值;

⑤ 双重循环进行两者比对, 如果 Item 描述属性中包含用户偏好项属性中的标签值, 那么 Item 与 User 相似;

⑥ 如果对应上的标签越多越相似.

3.3 项目推荐

Item 推荐功能根据 User 的原始数据、User 的行为数据, 为 User 推荐他们感兴趣的 Item, 该功能主要由 User 自定义 Item 推荐规则、偏好度计算、Item 过滤、Item 推荐与展示等过程组成, 下面详细介绍各个功能过程.

(1) User 自定义 Item 推荐规则: 初始化用户的时候, 系统会默认为 User 初始化一种 Item 推荐规则. 但是 User 也可以根据自己的需要定义 Item 推荐规则. 如原始信息匹配(角色、岗位)、User 相似度匹配、User 和 Item 偏好度匹配、Item 关联度匹配、行为数据分析的优先匹配顺序, 也可以进行勾选不进行该项匹配规则匹配. User 原始匹配规则是必须要保留的, 而且是最优先匹配.

(2) 偏好度计算: 根据 User 相似信息、User 与 Item 相似性、Item 关联性的值进行相似度高低排序. 该相似高低为 Item 推荐的主要依据, 相似度算法如下:

① 读取当前 User 的相似 User;

② 通过相似 User 获取最近访问的项目集合 ItemList1;

③ 读取当前 User 相似的 Item 集合 ItemList2;

④ 合并 ItemList1 和 ItemList2 形成一个新的 Item 集合 ItemListAll;

⑤ 循环 ItemListAll 拿其中的任意一个 Item 找到与它关联的 Item, 添加到 ItemListAll 中;

⑥ 根据 User 对 Item 的偏好高低、Item 评分高低进行 User 偏好度的计算, 根据某一特定偏好阀值进行取舍.

(3) Item 过滤: 主要是将 User 访问过的、根据法规等不适合推荐的 Item 过滤掉.

(4) Item 推荐与展示主要根据 Item 的偏好度和 Item 过滤后的推荐项进行选择的推荐和展示, 具体过程:

① 获取 User 设置的推荐个数 x;

② 按照相似性由高到低获取 x 个推荐项 Items, 将其推荐给 User.

推荐模型根据上述描述的过程获取推荐项, 并根据推荐项为用户提供个性服务, 基于标签的多因素的用户群与资源集关系如图 5 所示, 根据用户的多因素及各因素优先级将用户归类到用户群中, 根据该用户的偏好资源集查找到类似的资源集合推荐给该用户所在的用户群. 基于标签的多因素推荐提高了用户使用的体验度, 节省了用户的查询时间, 满足了用户的个性化需求. 具有准确、高效、个性化的特点.

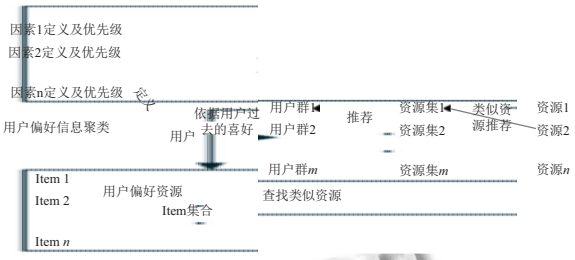

图 5 基于标签的多因素的用户群与资源集关系

4 结束语

采用基于推荐算法和标签的多因素推荐算法, 构建的基于标签的多因素用户群与资源集关系, 已应用于个性学习模式的远程培训平台, 该平台上有供学习者学习的各种资源 628 个, 共有文档、视频、答疑、通知、提醒、反馈、数据录入规范等 7 种资源, 面向 8 个工种及岗位, 提供给 1287 个用户同时使用. 平台运行初期使用的是基于内容的推荐和基于规则的推荐算法, 未使用协同过滤推荐是因为冷开始问题, 平台运行 3 个月, 采集数据进行分析, 推荐效果不加, 用户使用率下降 30%. 继而研究基于标签的多因素推荐算法并实施, 用户使用率得到了提升, 经一年多的实际应用测试, 为用户推荐的资源占用户所需资源的 90%, 用户所属群划分的准确率 89%, 资源分类的准确率 92%, 资源获取的平时等待时间 0.5s, 满足了学习者的需求, 提高了培训平台的使用效用.

参考文献

1 协同过滤. http://baike.haosou.com/doc/6971257.html.
2 刘志勇, 刘磊, 刘萍萍, 等. 一种基于语义网的个性化学习资源推荐算法. 吉林大学学报(工学版), 2009, 39(S2): 391–395.
3 冯亚丽, 孙龙安, 伊三泉, 等. 关于油田钻井公司信息化培训中现代远程教育模式的探索与实践. 成人教育, 2015, (2): 65–68.
4 主要的推荐算法简介. http://blog.sina.com.cn/s/blog_602feaa80100fjq9.html. [2009-11-08].
5 Goossen F, Jntema WI, Frasincar F, *et al*. News personalization using the CF-IDF semantic recommender. http://repub.eur.nl/pub/31385/. [2011-07-26].
6 Pazzani MJ, Billsus D. Content-based recommendation systems. The Adaptive Web. 2007. 325-341.
7 曾春, 邢春晓, 周立柱. 个性化服务技术综述. 软件学报, 2002, 13(10): 1952–1961.
8 张羽, 李越. 基于 MOOCs 大数据的学习分析和教育测量介绍. 清华大学教育研究, 2013, 34(4): 22–26.