文章编号：2095-2163(2021)08-0118-04 中图分类号：TP391 文献标志码：A

基于标签的在线学习资源推荐算法

文 谧[1]，朱木清[2]
（1 广州应用科技学院 广州，511370；2 广东工业大学华立学院，广州 511325）

摘 要：在线学习是目前获取知识的一种重要途径，然而信息过载导致从在线学习平台的大量资源中找到所需的学习资源非常困难。本文提出了一种基于标签的推荐算法，混合基于内容推荐和协同过滤推荐，采用 TF-IDF 来平衡热门标签的权重，采用修正的余弦函数相似性计算用户间、资源间的相似性，结合学科知识图谱，让推荐结果在相似基础上增加扩展性，满足进阶学习特点。实验结果表明，本文提出的算法在准确率和推荐效率上优于传统的协同过滤推荐算法，为解决同类问题提供了较强的参考价值。

关键词：标签；推荐系统；协同过滤；在线学习资源

Online learning resource recommendation algorithm based on tag

WEN Mi[1], ZHU Muqing[2]
(1 Guangzhou Institute of Applied Science and Technology, Guangzhou 511370, China;
2 Huali College, Guangdong University of Technology, Guangzhou 511325, China)

【Abstract】 Online learning is an important way to acquire knowledge. However, the overload of information makes it very difficult to find satisfactory learning resources from a large number of resources of online learning platform. This paper proposes a tag based recommendation algorithm, which combines content-based recommendation and collaborative filtering recommendation. TF-IDF is used to balance the weight of hot tags. The modified cosine function similarity is used to calculate the similarity between users and resources. Combined with the subject knowledge map, the recommendation results can be expanded on the basis of similarity to meet the characteristics of advanced learning. Experimental results show that the proposed algorithm is superior to the traditional collaborative filtering recommendation algorithm in accuracy and recommendation efficiency, which provides a strong reference for solving similar problems.

【Key words】 Label; Recommendation system; Collaborative filtering; Online Learning Resources

0 引 言

在线学习为人们提供了一个灵活、便捷的学习方式，是传统课堂学习之外获取知识的最重要途径。在线学习平台可以给用户提供优质的教育资源，满足用户个性化学习的需求，深受广大学习者欢迎。然而，随着在线学习课程资源的不断丰富，互联网学习资源信息过载问题，导致用户从在线学习平台的大量资源中找到需要的、满意的学习资源非常困难。

目前，使用搜索引擎可以解决部分信息过载需求，但搜索结果较局限、无区别化。推荐系统通过分析和计算用户的兴趣模型，发现用户难以表达的需求，更好满足用户全面的需求。可较好地解决如何准确提供在线学习资源的问题。

用户可以用标签标注自己感兴趣、关注或需求的学习资源[1]。这些标签作为用户的元数据，为个性化推荐系统提供了十分重要的数据基础。

1 基于标签的资源推荐算法

用户与学习资源通过标签建立强联系，基于标签的推荐可以更加个性化[2]。同时，针对推荐结果仅考虑相似性问题，提供结合知识图谱技术进一步拓展推荐结果，以满足进阶等学习特点。

1.1 推荐算法基本思想

基于标签的学习资源推荐算法分为 5 个环节：

(1)根据用户标签建立用户模型，计算用户间学习偏好相似性并修正，找到最近的用户邻居。

(2)将最近邻居的标签和用户的标签对比，通过计算获得与最近邻居相关度大并且与用户相关度小的价值标签。

(3)计算被价值标签标注的权重高且与用户暂无相关的资源，得到第一阶段推荐结果。

基金项目：广东高校省级重点平台和重大科研项目(青年创新人才类)(2016KQNCX212)。

作者简介：文 谧(1982-)，女，硕士，讲师，主要研究方向：计算机应用技术、人工智能；朱木清(1982-)，男，硕士，讲师，主要研究方向：智能信息处理、知识工程。

收稿日期：2021-05-24

(4)采用基于内容的过滤方法,根据用户学习偏好标签,对第一阶段推荐结果进行再次过滤,过滤掉无价值的学习资源,得到第二阶段推荐结果。

(5)根据知识图谱的关系,计算与第二阶段推荐结果最相关的知识资源,两者混合作为最终推荐结果。

推荐模型如图 1 所示。

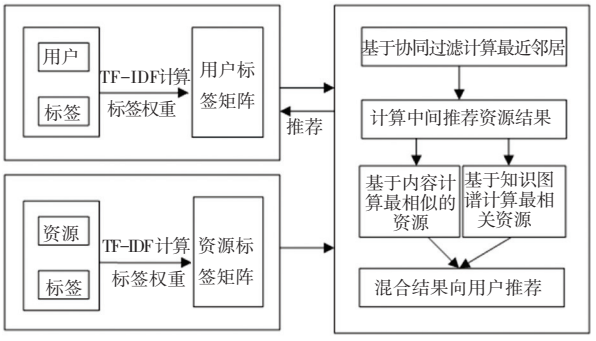

图 1　基于标签的推荐系统模型

Fig. 1　Recommendation system model based on tags

1.2　用户建模

在线学习资源推荐系统,要能够向用户提供个性化、准确和高效的推荐结果,首先要获取用户全面的资源标注信息,建立一个准确的用户模型。用户模型的准确性主要包括两方面:准确描述用户需求偏好和区分用户[3]。推荐系统将根据用户模型,更好地完成推荐任务。

标签是用户附加在资源上的关键词,利用 TF-IDF 度量每个标签,再进行用户相似性计算,可以平衡热门标签的权重,提高推荐结果的新颖性。

标签 T 对用户 U 的权重用 $W(T,U)$ 表示,采用 TF-IDF 平衡热门标签的权重。

$$W(T,U) = TF(T,U) \times IDF(T), \tag{1}$$

$$TF(T,U) = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{2}$$

$$IDF(T) = \log_2 \frac{N+1}{N(T)+1} + 1, \tag{3}$$

其中,$n_{i,j}$ 是标签 T_i 在用户 U_j 中的使用次数,分母则是在用户 U_j 中所有标签的使用次数和。计算每个标签对用户的权重可以得到用户与标签的权重数据。

$$N_{m\times n} = \begin{bmatrix} w(t_1,u_1) & w(t_2,u_1) & \cdots & w(t_n,u_1) \\ w(t_1,u_2) & w(t_2,u_2) & \cdots & w(t_n,u_1) \\ \vdots & \vdots & \vdots & \vdots \\ w(t_1,u_m) & w(t_2,u_m) & \cdots & w(t_n,u_m) \end{bmatrix}. \tag{4}$$

1.3　推荐对象建模模块

标签是用来描述信息的关键词,其特点为无层次结构。学习资源等对象模型和用户的需求偏好模型都可以采用标签的向量空间表示进行描述。标签对学习资源的重要性也采用 TF-IDF 表示,相应模型如下所示。

$$R_{m\times n} = \begin{bmatrix} w(t_1,r_1) & w(t_2,r_1) & \cdots & w(t_n,r_1) \\ w(t_1,r_2) & w(t_2,r_2) & \cdots & w(t_n,r_2) \\ \vdots & \vdots & \vdots & \vdots \\ w(t_1,r_m) & w(t_2,r_m) & \cdots & w(t_n,r_m) \end{bmatrix}. \tag{5}$$

1.4　推荐算法模块

1.4.1　协同过滤推荐

完成用户建模和推荐资源建模后,就可以对用户间的相似程度,以及学习资源间的相似程度进行计算。采取与用户对推荐对象的标签平均权重差值方法,可以在一定程度上解决原方法中不同用户可能有不同的标签权重问题。用 $I_{i,j}$ 表示用户 i 和用户 j 共同使用的标签集合。I_i 表示用户 i 使用的标签集合,I_j 表示用户 j 使用的标签集合,则用户 i 和用户 j 之间的相似程度 $sim(i,j)$ 如式(6)所示:

$$sim(i,j) = \frac{\sum_{t\in I_{i,j}} (W_{t,i} - \bar{W}_i)(W_{t,j} - \bar{W}_j)}{\sqrt{\sum_{t\in I_i} (W_{t,i} - \bar{W}_i)^2} \sqrt{\sum_{t\in I_j} (W_{t,j} - \bar{W}_j)^2}}, \tag{6}$$

其中,$W_{t,i}$ 表示用户 i 使用的标签权重,$\bar{W}_i$、$\bar{W}_j$ 分别表表示用户 i 与用户 j 使用的标签平均权重。

通过相似度度量方法计算出用户之间的相似性后,可从用户的邻居中选择与用户相似性最高的那些邻居。设用户 u 的最近邻居集合 U^*,记为:

$$U^*(u,t) = \arg\max_{u\in U} sim(i,j). \tag{7}$$

1.4.2　基于内容的推荐

根据用户学习偏好标签,采用基于内容的推荐方法,对第一阶段推荐结果进行过滤,过滤掉用户不感兴趣的学习资源,得到中间推荐结果。

$$R(u,r) = \operatorname{argmax}_{r\in R} \sum_{u\in U^*} sim(i,j) \times w(t,u) \times w(t,r), \tag{8}$$

过滤用户不需要的学习资源方法,是在建模基础上,计算中间推荐资源与用户学习标签向量的相似性,得到第二阶段推荐结果。计算公式如下:

$$R^*(u,r) = \operatorname{argmax}_{r\in R} sim(i,j). \tag{9}$$

其中,$sim(i,j)$ 是计算中间推荐结果的标签向

量与用户的兴趣资源标签向量相似性的函数,而 R^* 则是基于相似性的推荐结果。

1.4.3 混合推荐策略

在相似性推荐结果的基础上,根据学科领域知识图谱的边权重大小,计算与相似性推荐结果资源最相关的知识资源,混合两者作为最终推荐结果。例如:计算机学科课程知识图谱的示意如图 2 所示[4]。

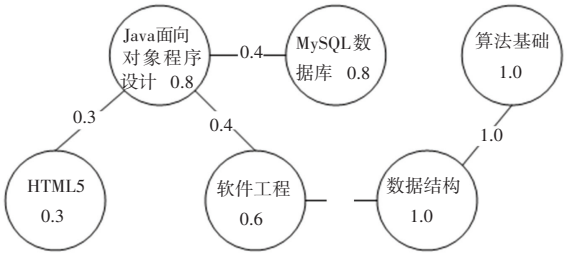

图 2 计算机领域知识图谱实例

Fig. 2 Examples of knowledge graphs in the computer domain

在图 2 中,“数据结构”与“算法基础”的边权重相比“软件工程”与“数据结构”的边权重大,说明“数据结构”与“算法基础”的相关度更高。对“数据结构”学习资源感兴趣的用户,很可能也需要最相关的“算法基础”学习资源。因此,将最相关的资源和 R^* 混合推荐给用户,使得推荐结果既有相似性又有扩展性,更好满足用户需求。

2 实验结果与分析

2.1 实验数据

本文实验采用的数据集为 Goodbooks-10k 和 Delicious。首先,对于数据集的噪声数据进行预处理,删除数据不完全的记录和标签数较少的用户数据,保证实验数据的合理性。

2.2 实验设计

基于以上实验数据,验证本文提出的基于标签的混合推荐算法是否更有效,结果对比协同过滤推荐算法。通过平均绝对误差、平均平方误差和标准平均误差等典型指标对准确度进行衡量[5]。

推荐资源对象列表中,用户需要的标签与系统中用户标注的所有标签的比例,计算方法如式(10)所示。

$$R = \frac{N_{rs}}{N_r}. \tag{10}$$

推荐资源对象列表中,用户需要的标签和所有被推荐标签的比例,计算方法如式(11)所示。

$$P = \frac{N_{rs}}{N_s}, \tag{11}$$

基于上述两项计算,进而计算推荐效率.计算方法如式(12)所示。

$$F = \frac{2PR}{P + R}. \tag{12}$$

其中,N_{rs} 为推荐列表中用户需要的标签个数;N_r 为用户标注的所有标签的个数;N_s 为所有被推荐标签的个数。

2.3 实验结果分析

对两个推荐算法进行了运算,并计算其准确率以及推荐效率。采用本文提出的方法,基于标签进行系列计算后,再根据知识图谱的关系计算与相似性推荐结果最相关的知识资源,混合两者作为最终推荐结果。两个推荐算法的准确率和推荐效率见表 1,数据对应曲线分别如图 3、图 4 所示。

表 1 不同推荐列表长度的数据参数

Tab. 1 Data parameters of different recommended list lengths

TopN	基于标签的混合推荐算法		协同过滤推荐算法	
	precision	recall	precision	recall
10	0.119 3	0.392 5	0.109 1	0.280 3
20	0.142 5	0.413 7	0.127 6	0.305 7
30	0.159 7	0.484 9	0.138 7	0.351 4
40	0.162 3	0.602 1	0.149 1	0.440 3
50	0.179 3	0.664 5	0.156 2	0.485 8
60	0.187 4	0.756 3	0.160 7	0.611 2
70	0.193 8	0.816 5	0.163 1	0.709 3
80	0.198 5	0.857 3	0.161 3	0.743 2

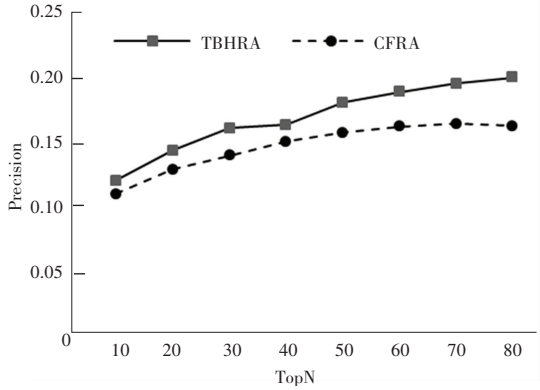

图 3 推荐算法的 precision 比较

Fig. 3 Comparison of precision of recommended algorithms

由实验结果可以看出,基于标签的混合推荐算法具有较高的准确率和召回率。由于进行了平衡权重并修正,且结合知识图谱完善推荐资源,使得计算结果更加准确和全面。

(下转第 125 页)