在线学习资源混合推荐算法研究

张群慧　朱爱军
（湖南科技职业学院，湖南 长沙 410004）

摘　要　在线学习网站上为增加对用户的满意度和忠诚度会使用推荐算法，来向用户介绍其所喜欢的学习资料。因此，推荐算法的精准度是一个首要问题。目前，最主要的推荐方式有三类：基于内容的推荐算法、协同过滤算法及其混合推荐算法。基于内容的推荐算法与协同过滤算法各有弊端。而且基于内容的推荐算法还具有与提供相关内容信息交互的弊端，而协同过滤算法有数据稀疏性和扩展性方面的问题。在这两个方法的基础上，混合的推荐算法取长补短，把两个算法的优点组合起来，为人们提供更好的服务。

关键词　在线网站　学习资源　混合推荐
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1007-0745(2022)11-0001-06

随着网络科技的日益发达和社会物质生活的日益改善，为了在这个快节奏的社会节省不必要的时间以及提高生活质量，个性化推荐成为互联网的核心功能之一，并被运用于各个行业。在线学习网站使用推荐算法，为用户推荐最有机会学习和购买的课程资料，给用户带来潜在的好友关注；为用户推送其最有机会点击的视频内容[1]；为用户提供最有趣的特色信息。个性化推送技巧是解决之道，这也是互联网智能的体现之一。

现阶段的选择方式大致有三类：基于内容的推荐算法、协同过滤算法、混合推荐算法。[2]

基于内容的推荐算法根据具体的文本数据来进行选择。

协同过滤算法一般包含了通过程序的计算和通过模式的计算。[3]在最初的时候一般是通过大数据研究，利用问题的评估矩阵的实验数据，来寻找类似的问题或者数据。然后通过构建一个模式，以通过这个模式的评估矩阵为基础实验数据，来进行智能的推荐预测。

不过，上面的两个推荐算法都有着干扰最终选择结果的显著缺陷，于是科学家们为了避免两个算法的弊端，发挥其长处，给出了第三个算法，混合推荐算法，它是一个把基于内容的推荐算法与协同过滤算法融合到一起的算法。现在大部分市场上的推荐系统都是通过各种推荐算法融合的混合推荐系统。[4]

1 协同过滤算法

1.1 基本原理

协同过滤算法是从海量的数据中挖掘出与用户兴趣相似的用户，使他们成为用户的邻居，然后根据他们喜欢的东西，产生一个推荐列表推送给用户。由此，可以看出协同过滤算法有两个核心问题：（1）确定与当前用户兴趣相似的用户；（2）产生推荐列表推送给当前用户。

要实现协同过滤算法有三大步骤：（1）收集用户的喜欢度；（2）找到相似的用户和学习资源；（3）计算数据并生成推荐列表。

1.2 协同过滤算法的问题

数据稀疏性问题也是协同过滤算法中必然会遇到的最大的问题。[5]在现实的商业推荐系统中，虽然用户以及与其相应的购物项目的信息总量都非常巨大，但用户却通常只能在很少的购物项目上基于客观评价，从而产生评价矩阵数量稀少的问题。在数量如此稀少的情形下，由于没有参考的数据信息，不利于使用所选择的机器学习资源估计两个不同用户间的接近程度，进而造成对邻居集合的资源选取不精确，影响了推荐精度。冷启动也是关于数量稀少性的一种很典型的问题。所谓冷启动，即在没有用户评价的情况下，很难凭空生成符合当前用户需求的推荐列表。

扩展性问题是随着系统的使用所出现的问题。随

★基金项目：湖南省职业教育教学改革研究项目《人工智能驱动下太极环模型高职学生自适应学习系统的研究与实践》（ZJGB2019076）。

表 1 基于内容的推荐算法和协同过滤算法的比较

推荐算法	优点	缺点
基于内容的推荐算法	推荐结果直观，没有冷启动问题	用户和学习资源之间的关联模型构建需要大量的时间和人力；新用户问题；稀疏性问题
协同过滤算法	实现跨兴趣推荐，自动化程度较高	冷启动；数据稀疏性问题；扩展性问题

着用户和对应的购买学习资源的增多，由于计算资源和计算速度的限制，协同过滤算法在用户和对应购买学习资源增长到一定数量后，效率会大大降低，以致于不能满足实际需求。

2 基于内容的推荐算法

2.1 基本原理

基于内容的推荐算法的核心是学习资源之间两两相似度的精确度。如何实现这个算法，可分三步：（1）获取过去机器学习数据的特征值；（2）通过特征数据来构建用户的喜好模式；（3）通过将候选学习资源与喜好模型进行比较，生成推荐列表。

2.2 基于内容推荐算法的问题

推荐对象的多样性问题是基于内容的推荐算法的首要问题。喜好模型是根据过去用户喜欢的学习资源的学习中得出的，所以在原始数据中就有隐患。即过去喜欢的学习资源不可能包括用户潜在喜欢的学习资源。

同时对用户和学习资源间建立关联模型和提取主要特征是一个会耗费大量的时间和人力的过程。此外，如果用户很少有购买和浏览的行为，那么所得到的数据总量就远远不够，将会大大地影响推荐结果的精准度。

3 混合推荐算法

3.1 基本原理

协同过滤算法与基于内容的推荐方式是能够优势互补的。协同过滤算法并不出现推荐学习中资源多样性的现象，而基于内容的推荐方式并不出现资源冷启动的现象。因此可以将这两种推荐算法融合起来成为一种新的混合推荐算法，以求更好的推荐效果。

3.2 算法种类及优势

该方案将改变传统基于内容的方式获取用户既有兴趣，而采用特征词的协同过滤方式获取用户潜在兴趣，并将用户现有的兴趣和潜在兴趣混合，得出混合的用户兴趣模型，用混合模式与候选学习资料进行相似度统计，给不同用户推送可能感兴趣的学习资料。较之前的方法，本文更充分地考虑了用户喜好以及在多样性与个性化上的要求，更加充分地挖掘了用户的潜在兴趣，提高了用户对推荐学习资源的点击率。

4 在线学习资源混合推荐算法

混合推荐系统模型主要由三个部分组成：用户已有的兴趣模型、潜在的用户兴趣模型、混合推荐算法模型。首先在这里有些定义需要说明一下。

特征词序列：搜索内容有非结构化和结构化之分。结构化是指搜索内容就是一个或多个确定的词汇，例如：牛津字典、书包等。非结构化内容指的是用一个长短语或句子来描述想要检索的学习资源信息。推荐算法是以结构化的搜索内容为基础的，所以要对搜索内容的文本信息进行结构化处理，这一部分的实际操作就是提取搜索内容的特征词。对任意用户的搜索非结构化内容的集合 D={d1,d2,d3.....dn}，将能够代表搜索内容和搜索内容特征的词汇或者短语通过数学算法提取出来形成一个特征词序列 S={s1,s2,s3.....sn}。特征词序列包含了用户的非结构化内容集合的特征词，类似于给每个用户搜索内容集贴上标签。

用户已有的兴趣模型（EM）：将任意用户的搜索内容的文本信息向量化，然后经过一些数学运算，算出每个特征值的权重，得出特征词序列的一一对应的特征值权重向量，记为 W1={w11,w12,w13,...,w1n}，称其为用户已有的兴趣模型。这是根据搜索内容来设计兴趣模型，这属于基于内容的推荐算法的一部分。

用户潜在的兴趣模型（PM）：通过协同过滤算法找出当前用户的邻居，邻居就是与当前用户兴趣相似的用户群体，将邻居的已有兴趣推荐给当前用户，即将相似的用户的已有兴趣模型作为当前用户的潜在兴趣模型，记为 W2={w21,w22,w23,...,w2n}，其中 W2i 是特征词序列中对应的权重。

混合兴趣模型（HM）：将用户已有兴趣模型和潜在的兴趣模型按照一定的规则合并得到的权重向量，记为 W3={w31,w32,w33,...,w3n}，其中 W3i 是特征词序列中对应的权重，称其为混合兴趣模型。

混合兴趣模型中用户已有兴趣模型根据原理可划分为基于内容的推荐算法的版块 [6]，用户潜在兴趣模型属于协同过滤算法的内容，所以这就是混合兴趣推荐

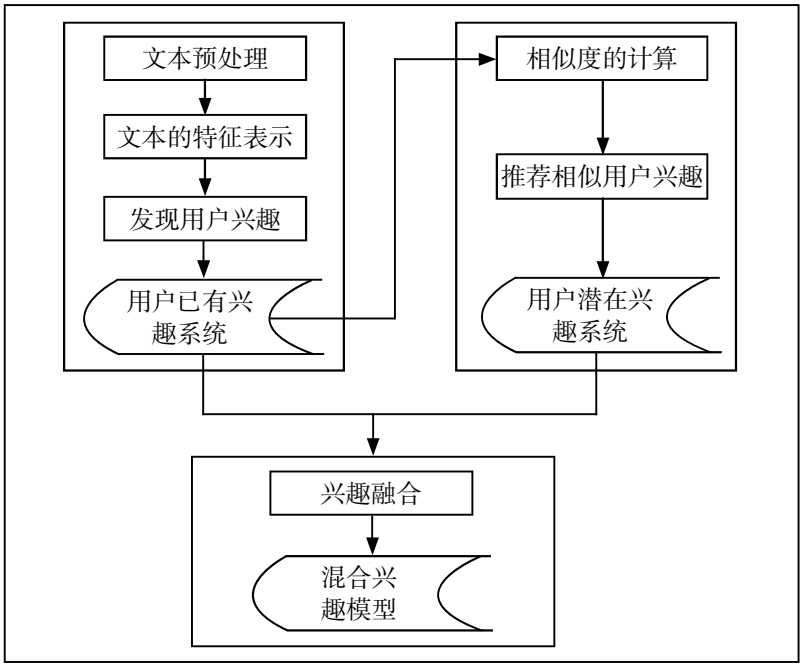


图 1 混合兴趣模型构建流程

模型的基本框架。

4.1 用户已有的兴趣模型设计

建立用户的兴趣模型之前要对搜索的内容进行结构化处理。典型的处理办法就是 TF_IDF（term frequency_inverse document frequency）表示法。这个表示法是用权重来衡量词汇的重要程度。TF（term frequency）词频，即某个特定的词汇在文本信息中所有词汇中所占的比例。在所有的文本内容里，“的”“这些”“那些”，类似这种没有实意的常见词的词频通常会很高，因此为了提高提取的特征词的准确度，提出了 IDF（inverse document frequency）逆文档频率的概念。假设某个词比较少见，却在某个文档出现次数较多，那么这些词汇很有可能反映文章的特性，极有可能是文章内容的特征词。为这种情况设置了一个新的权重参数便是 IDF，这个参数和词汇的常见程度成反比。TF_IDF 的权值计算方法为 TF*IDF，[freq(i,j)/sum(k,j)]*log[N/n(j)]，TF=freq(i,j)/sum(k,j)，其中 freq(i,j) 为在搜索内容集 dj 中词汇 i 出现的次数；sum(k,j) 为在搜索内容集 dj 中所有的词汇个数。IDF=log[N/n(j)]，其中 N 为搜索内容集的总数，n(j) 为出现过词汇 i 的搜索内容的条数。

给定搜索内容集 D={d1,d2,...,dn}，和特征词集合 S={s1,s2,...,sn}，搜索内容集可表示为与特征词集合 S 对应的一个向量空间模型，di={wi1,wi2,...,wij,...,wik}，其中 wij 表示特征词 sj 在搜索内容集 di 的权值，如果 wij 为 0，表示为在搜索内容集 di 中没有特征词 sj，于是，搜索内容集可以等同于一个权值矩阵：

$$DM=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & \cdots & w_{2k} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nk} \end{bmatrix}$$

4.2 用户潜在的兴趣模型设计

用户的潜在兴趣模型与用户的已有兴趣模型的区别在于不能通过以往的搜索内容直接发现用户潜在的兴趣。本文提出用协同过滤算法来解决问题。传统的协同过滤算法是通过评分矩阵来发现相似用户，通常不同的用户会购买或者浏览相同的学习资源，但是购买不同的学习资源，这些浏览了相同的学习资源的用户很难归为相似用户。针对上述问题，我们只要计算不同用户搜索内容的相似度 sim(u,v) 即可。

4.2.1 相似度的计算

协同过滤算法的核心部分是寻找兴趣相似的用户，其效率和结果很大程度上决定了协同过滤算法的效率和结果。测度用户 i 和用户 j 的搜索内容相似性的方法如下：首先得到用户 i 和 j 的搜索内容集特征词权重的所有项，然后通过相似度测量方法来计算用户 i 和 j 的相似度，记为 $sim(i,j)$。

通常 $sim(i,j)$ 的计算方法有三种：余弦相似性计算

表 2 数据表示形式

	$Term_1$	……	$Term_k$	……	$Term_n$
$Content_1$	W_{11}	……	W_{1k}	……	W_{1n}
……	……	……	……	……	……
$Content_j$	W_{j1}	……	W_{jk}	……	W_{jn}
……	……	……	……	……	……
$Content_m$	W_{m1}	……	W_{mk}	……	W_{mn}

表 3 推荐结果

	推荐	未推荐
访问	推荐该学习资源并被访问了（rv）	未推荐该学习资源却被访问了（nv）
未访问	推荐了未被访问（rn）	未推荐该学习资源也未被访问（nn）

法、皮尔逊相关系数、欧几里得度量，本文采用余弦相似性计算法。

设用户 u 搜索内容集 Du={du1,du2,...,dui,...,dum},EMu=(w1u1，w1u2，…，w1uj，…，w1uk)，用户 v 搜索内容集 Dv={dv1，dv2，…，dvj，…，dvs}，Du、Dv 均为 D 的子集，EMv=(w1v1，w1v2，…，w1vj，…，w1vk)，用户 u,v 的内容相似度如下：

$$sim(u,v)=\cos(u,v)=\frac{EM_u \bullet EM_v}{|EM_u|\times|EM_v|}$$

4.2.2 推荐相似用户群的兴趣词并构建模型

通过上述算法可以计算出当前用户与其他所有用户搜索内容之间的相似度，排列出与当前用户相似度最高的 n 个用户，作为邻居群。用协同过滤算法将邻居群的用户的已有兴趣模型推荐给当前用户，即为用户的潜在兴趣模型。

4.3 混合推荐算法模型的设计

得到用户已有和潜在兴趣模型后，将两个兴趣模型按照规则合并，再与候选的推荐学习资源集合计算相似度[7]，给定相似度阈值 a，检查推荐结果。

设用户 u 的 EMu=(w1u1，w1u2，…，w1uj，…，w1uk)，PMu=(w2u1，w2u2，…，w2uj，…，w2uk)，HMu=(w3u1，w3u2，…，w3uj，…，w3uk)，候选推荐内容集 d=(wd1，wd2，…，wdj，…，wdk)。w3 的计算方法为：

W3= max{w1,w2}，其中 max{} 表示 w1,w2 中的较大值。最后用余弦法计算 d 和 HM 的相似度，检查推荐结果。

算法：HM 的构建算法和推荐结果的生成，如下：

输入：用户 u 的 EMu、PMu，候选的搜索内容集 d，阈值 a。

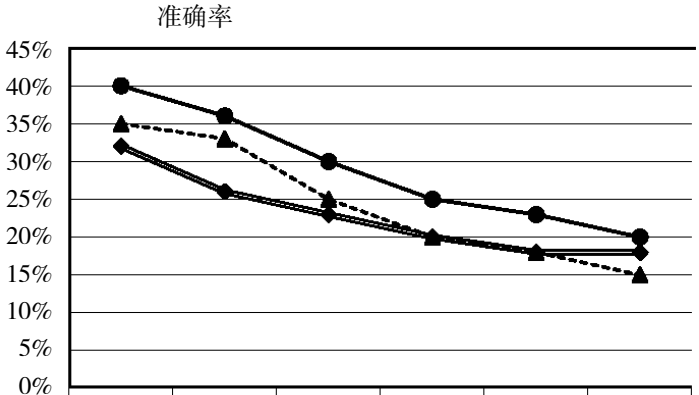
输出：候选搜索内容集 d 对用户 u 的推荐结果。

```
Begin
multi =0,moduleu = 0,moduled = 0,isRecommend = 0;
HMu =（w3u1,w3u2,...,w3uj,...,w3uk）;
for each sj ∈ S
w3uj =max（w1uj,w2uj）;
end for;
for each Sj ∈ S
multi =W1uj*W1vj;
moduleu =W1uj * W1uj;
modulev=W1vj*W1vj;
end for;
sim(u，v) = multi/(squr(moduleu)* squr(moduled));
if sim(u,d)> a
isRecommend =1;
end if;
End
```

5 实验结果及分析

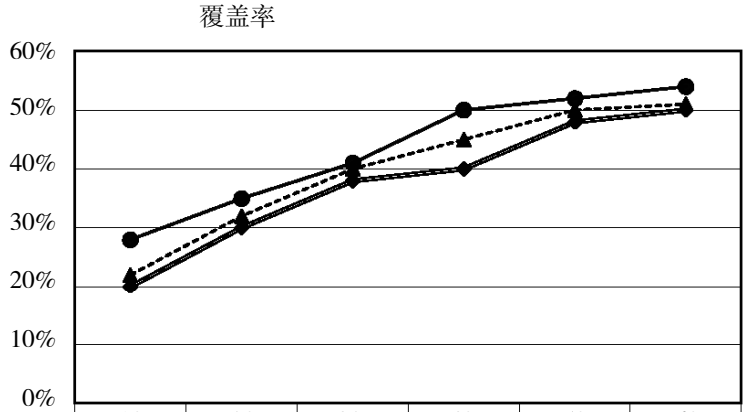
5.1 数据表示

在本文介绍的混合推荐系统中，是根据不同的搜索内容集中的各个特征词的不同权重来产生推荐结果的。权重矩阵可以用一个 m×n 的矩阵表示。m 行代表 m 个用户的搜索内容集，n 列代表 n 个特征词，第 i 行



	10	20	30	40	50	60
基于内容的推荐算法	35%	33%	25%	20%	18%	15%
协同过滤算法	32%	26%	23%	20%	18%	18%
混合推荐算法	40%	36%	30%	25%	23%	20%

图 2 准确率



	10	20	30	40	50	60
基于内容的推荐算法	22%	32%	40%	45%	50%	51%
协同过滤算法	20%	30%	38%	40%	48%	50%
混合推荐算法	28%	35%	41%	50%	52%	54%

图 3 覆盖率

第 j 列元素 Wij 表示第 j 个特征词在第 i 个用户的搜索内容集中的权重。权重矩阵如表 2 所示。

5.2 实验的评价指标

将推荐结果推荐给用户后有如表 3 结果：

根据结果的这几种可能，通常用准确率（或查准率，Precision）和覆盖率（或召回率，Recall）来作为算法评价的指标。准确率的计算公式为：

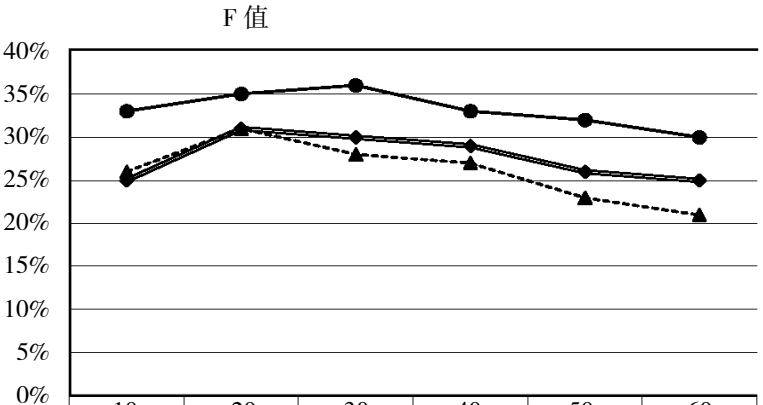
$$precision = \frac{rv}{rv+rn}$$

表示推荐了并被访问了的学习资源数量与推荐学习资源总数之比，而覆盖率计算公式为：

$$\mathrm{Re}call = \frac{rv}{rv+nv}$$

表示推荐命中学习资源数量与测试集中用户所访问学习资源总数之比。

实际上查准率和查全率是相互冲突的。如果增大推荐学习资源的数目，就会使得覆盖率增大，但是同时又使得准确率下降。因此，通常将两者给一个相当的权重合并成一个综合测度 F 来评价推荐质量。F 值越大，推荐质量越高。计算公式如下：



	10	20	30	40	50	60
基于内容的推荐算法	26%	31%	28%	27%	23%	21%
协同过滤算法	25%	31%	30%	29%	26%	25%
混合推荐算法	33%	35%	36%	33%	32%	30%

图 4 F 值

$$F = \frac{2 \times \mathrm{R}ecall \times precision}{\mathrm{R}ecall + precision}$$

5.3 实验方案

本文的数据取自 Datacastle 的用户浏览数据集，随机抽取 1000 名用户，将推荐项目数从 10 到 60 进行试验。发现用户潜在兴趣的过程中，要实现确定邻居群的大小即兴趣相似的用户个数。为了便于评估性能，我们把邻居群固定大小 35 人，相似度算法选用余弦计算法。推荐算法的最终结果是要生成 N 项用户可能最有兴趣的学习资源，以此供用户进行选择，考察在不同的 N 值下，不同的算法（基于内容推荐算法、协同过滤算法、本文介绍算法）的准确率、覆盖率以及 F 值。

由此，在不同推荐项目数（N）的情况下：

各个算法的准确率如图 2 所示。

各个算法的覆盖率如图 3 所示。

各个算法的 F 值如图 4 所示。

从上述的实验结果，我们可以得出以下结论：

（1）准确率和覆盖率是两个互逆的参数，随着推荐项目数的增大，准确率下降，覆盖率上升；（2）F 作为一个综合的参数权值，随着推荐项目数的增大有一个峰值，然后缓慢减少；（3）本文介绍的算法优于协同过滤算法和基于内容的推荐算法，同时还不存在冷启动的问题。

6 结论

推荐系统经历了较长时间的研究和发展已取得令人瞩目的成果。个性化推荐系统的作用主要表现在三个方面：（1）将在线学习资源的浏览者转变为购买者；（2）提高各种在线学习网站交叉融合的能力；（3）改善用户体验，提高用户忠诚度。

但是还需进一步的努力。不可否认的是，推荐系统还有许多难点没有突破，例如提取精准的用户偏好和对象特征；推荐的多维度研究；推荐系统的安全性研究等问题，但是我们坚信随着社会的发展，科技的不断进步，对于推荐系统的研究也会越来越深入，从而更好地服务于人们的物质文化生活。

参考文献：

[1] 尚松涛，石民勇，尚文倩，等．基于大数据的微视频推荐算法研究 [J]. 中国传媒大学学报（自然科学版），2017,24(02):38-45.

[2] 孙光浩，刘丹青，李梦云．个性化推荐算法综述 [J]. 软件，2017,38(07):70-78.

[3] 向小东，邱梓咸．基于 slope one 算法改进评分矩阵填充的协同过滤算法研究 [J/OL]. 计算机应用研究，2019(05):1-5.

[4] 杨海龙，李松林，李卫军．基于 GLSLIM 模型的混合推荐算法研究 [J]. 信息与电脑（理论版），2017(20):77-80.

[5] 杨丰瑞，郑云俊，张昌．结合概率矩阵分解的混合型推荐算法 [J]. 计算机应用，2018,38(03):644-649.

[6] 杨帅，王鹃．基于堆栈降噪自编码器改进的混合推荐算法 [J/OL]. 计算机应用，2018(07):1866-1871.

[7] 何慧．基于高斯模型和概率矩阵分解的混合推荐算法 [J]. 统计与决策，2018,34(03):84-86.