文章编号：2095-2163(2023)11-0197-05 中图分类号：TP391 文献标志码：A

基于新闻场景下的个性化推荐算法研究

王君威，余 粟
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：互联网下，每时每刻产生的新闻报道堪称海量，用户很难从海量报道中获取有用信息，推荐是解决该问题的重要方案。协同过滤推荐算法常用于物品推荐，在新闻场景下，新闻推荐与物品推荐有些许不同，本文在传统协同过滤的基础上针对新闻场景的特性提出了一种个性化新闻推荐方式，通过分析用户活跃度和新闻时效性对推荐的影响，对协同过滤做出改进、从物品和用户的角度进行新闻召回，使用冷启动进行召回补充，筛选召回结果，选取 Top$K$ 新闻报道产生个性化推荐结果。实验结果表明，本文针对新闻场景下设计的个性化推荐模型在数据稀疏和冷启动情况下推荐效果较优。

关键词：新闻推荐；个性化推荐；协同过滤

Research on personalized recommendation algorithm based on news scene

WANG Junwei, YU Su
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Under the Internet, the news reports generated at every moment can be called massive, and it is difficult for users to obtain useful information from the massive reports, and recommendation is an important solution to this problem. Collaborative filtering recommendation algorithm is often used in item recommendation, in the news scenario, news recommendation and item recommendation are slightly different. This paper proposes a personalized news recommendation method for the characteristics of news scene on the basis of traditional collaborative filtering, by analyzing the impact of user activity and news timeliness on recommendation, making improvements to collaborative filtering, news recall from the perspective of items and users, using cold start for recall supplementation, screening recall results, and selecting Top$K$ news reports to generate personalized recommendation results. Experimental results show that the personalized recommendation model designed in this paper for news scenarios has a good recommendation effect under the condition of sparse data and cold start.

Key words: news recommendations; personalized recommendations; collaborative filtering algorithm

0 引 言

随着网络的发展和电子设备的普及，人们已经不再只是通过电视的新报道和纸质报刊了解当下正在发生的事情，利用电子设备获取新闻消息成为当前的主流趋势。打开电子设备时，会看到太多对用户无用的嘈杂信息。对用户而言，没有好的推荐将很难获取有用信息，若不能根据用户个性化的需要加以筛选，信息过载的情况就难以避免。用户需求明确时，搜索能快速帮助用户满足需求。当用户需求不明确时，个性化新闻推荐成为了重要解决方案。

随着今日头条、抖音等 App 的飞速发展，推荐系统的价值已得到了业界的广泛认可。能够向用户提供有用信息，满足用户需求的新闻推荐也得到了相关学者的关注。万梅等学者[1]根据社交网络不断演化的特点，提出了一种基于神经网络嵌入和社交网络动态演化的新闻推荐算法，基于神经网络嵌入模型来学习用户在时间序列和社交网络中的行为特征，再使用无监督的随机游走学习用户的网络拓扑结构，通过插值法将行为特征向量和网络拓扑结构汇总表示用户向量，使用核函数将用户的交互映射到低维的向量空间，从而提高了相似度计算的效率。刘羽茜等学者[2]考虑到新闻的内容特征和浏览新闻之间的关系、相应的时序变化，以及不同新闻对用户的重要性等方面，提出了一种全面提取用户和新闻特征的新闻推荐模型。该模型基于深度学习，使用注入注意力机制的卷积神经网络深度提取新闻文本特征，通过对用户的新闻浏览记录进行时

作者简介：王君威(1997-)，男，硕士研究生，主要研究方向：推荐算法。Email：jwww925@ formail.com；余 粟(1962-)，女，博士，教授，主要研究方向：大数据、机电控制、计算机视觉。

收稿日期：2022-11-10

序处理,获取用户的兴趣趋势,使得注入注意力机制的深度特征融合新闻推荐模型角度更加全面。张丹等学者[3]考虑可以基于用户历史浏览的时间序列来做推荐,使用隐马尔可夫模型,并在其基础上加入了状态驻留的时间元素,将隐马尔可夫模型的五元素扩展为六元素,以此来提高新闻推荐的准确度。袁仁进等学者[4]针对新闻推荐中的用户画像构建与用户兴趣流动的问题,提出一种面向新闻推荐的用户兴趣模型构建和更新方法,使用向量空间与 bisecting K-means 一同构建原始用户兴趣模型,然后使用艾宾浩斯遗忘曲线为基础构造遗忘函数来描述用户的兴趣变迁,更好地反映用户的兴趣变化。陶天一等学者[5]将知识图谱中“实体-关系-实体”表示事物关系的形式引入到新闻推荐中来,提出 KHA-CNN,基于卷积神经网络和层次注意力机制获取的新闻报道的特征,根据用户的交互来获取用户画像,基于两方特征实现推荐。王利娥等学者[6]看到当前的推荐模型不能有效地平衡差分隐私的噪声,提出了一种融合知识图谱和隐私保护的新闻推荐方法,采用多通道卷积神经网络模型融合新闻标题、知识图谱中实体和实体上下文等多维度的特征向量;其次,利用注意力机制为不同敏感程度的特征向量添加不同程度的噪声,从而降低噪声对数据分析的影响;然后,对加权的用户特征向量添加统一的拉普拉斯噪声,以保证用户数据的安全性;利用两阶段的隐私保护机制来提高用户的隐私保护和推荐效果。朱文跃等学者[7]针对传统推荐系统中存在的共性问题,冷启动、数据稀疏、推荐精度较低等问题,提出了基于事件本体的推荐算法,结合新闻结构和语义构建事件本体,对用户浏览的新闻抽取要素,构建用户兴趣模型,综合事件本体的相似度、用户兴趣相似度、非层次结构相似度三个角度完成推荐。

本文分析了用户活跃度和新闻时效对新闻推荐的影响,针对新闻场景提出一种个性化推荐方式。通过新闻时效性和用户活跃度的判断, 对 *item* 和 *user* 的协同过滤进行改进、并加上冷启动进行新闻报道的召回,尽可能获取包含用户兴趣点的所有新闻报道,基于目标用户的评分进行排序,产生面向用户的个性化推荐。

1 相关工作

1.1 场景描述

新闻报道的推荐和传统的物品推荐有些许不同,在物品推荐中,物品的数量和种类在相对长的一段时间里面会保持稳定,不会发生较大的变化;而且短期内用户对物品的喜好也不会发生大的变动。与之不同的是,世界上每天都会发布大量的新闻报道,新闻的数量随着时间的推移也会越来越多。

在新闻推荐中已经报道的新闻,如果当时没有及时推荐给感兴趣的用户,那么在相对较短的一段时间之后,不会再推荐对应的用户。对用户来讲,一般情况下也不会对过往的新闻报道感兴趣。物品推荐往往还存在用户对推荐结果的显式评价(包括点赞、关注、购买、评价等行为),而新闻报道的推荐只能根据用户的历史交互来判断推荐结果。除此之外,活跃度高的用户交互记录较为丰富,而活跃度低的用户交互记录相对稀疏,在判断目标用户的兴趣相似集合时,往往会发现活跃度高的用户与绝大部分用户的兴趣都相近,难以区分用户的相似情况,这就对个性化新闻推荐提出了较高的要求。

1.2 相关模型

本文提出的个性化新闻推荐整体流程如图 1 所示。在新闻平台数据集的基础上,对用户交互、新闻报道等相关信息进行数据分析和处理,将数据集通过召回层,从 ItemCF 和 UserCF 以及冷启动的角度对新闻进行召回,并基于相似用户集合或新闻集合对召回新闻的评分进行排序,最后产生目标用户的个性化新闻推荐。

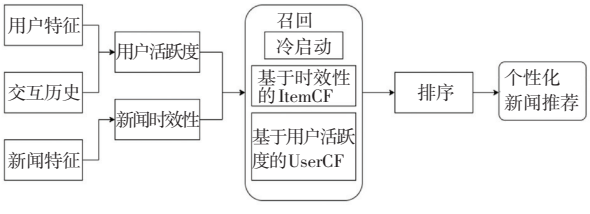

图 1 个性化新闻推荐模型

Fig. 1 Personalized news recommendation model

1.2.1 协同过滤

协同过滤(Collaborative Filtering)是推荐领域中最重要的算法,实质是根据用户的交互历史以及拥有类似兴趣的用户的交互记录或曾经感兴趣的物品来给目标用户进行物品推荐。主要分为基于 item 的协同过滤和基于 user 的协同过滤。2 个算法的思想基本一致,核心在于类似用户判定或喜好物品的相似物品判定,即物品或用户相似度的计算。

基于 item 的协同过滤是根据用户曾经喜欢的物品,从数据集中找到相似的物品,推荐给用户。而基于 User 的协同过滤是基于用户的交互历史,从用户集合中找到与目前用户的兴趣较为接近的用户子集,基于这些不同用户的交互喜好,给目标产生新的

推荐。

协同过滤根据用户与物品的历史交互，将用户点赞、收藏、转发等不同的行为进行量化，将量化后的行为作为用户对物品的评分，构建成评分矩阵，评分大小表示用户感兴趣的程度。无论基于 item 的协同过滤、还是 User 的协同过滤都是基于评分矩阵 $\boldsymbol{R}$ 找到与相似物品或相似用户，基于同样兴趣爱好的人群或相似喜好物品的集合产生的个性化推荐。

由于新闻推荐场景与物品推荐场景不同，在新闻场景下很少有用户看过新闻后会基于对新闻的喜好进行点赞、收藏、转发，显式行为数据的极度稀疏导致无法根据通用方式完成新闻评价，本文使用用户对新闻的历史交互点击作为用户对物品评分的依据。

以基于 User 的协同过滤算法为例，评分矩阵形式如下：

$$\boldsymbol{R}=\begin{bmatrix} u_{(1,1)} & \cdots & u_{(1,n)} \\ \vdots & \ddots & \vdots \\ u_{(m,1)} & \cdots & u_{(m,n)} \end{bmatrix} \tag{1}$$

其中，$\boldsymbol{R}$ 表示评分矩阵；U 表示用户集合；I 表示物品集合；m 和 n 分别表示用户和物品的个数；使用 $u_{(i,j)}$ 表示第 i 个用户对第 j 个商品的评分。

协同过滤算法常用余弦相似度或皮尔逊相关系数来衡量用户或物品的相似程度。

余弦相似度公式如下所示：

$$sim_{uv}=\frac{\sum_i r_{ui} * r_{vi}}{\sqrt{\sum_i r_{ui}^2}\sqrt{\sum_i r_{vi}^2}} \tag{2}$$

其中，u 与 v 分别表示不同用户的评分向量；i 表示物品序号；r_{ui} 为用户 u 对第 i 个物品的评分；$\sum_i r_{ui}$ 表示用户 u 对物品集合中所有物品的评分集合中的第 i 个物品的评分向量。

在现实情况下由于不同用户对物品评分的偏好不同，有的习惯给高分、有的经常给低分，会造成评价不统一，影响推荐的效果，为了平衡不同用户对物品的评分偏差，皮尔逊相关系数以不同用户的平均分为平衡点对评分进行修正，降低评分的偏差。

皮尔逊相关系数计算方式如下：

$$sim(u,v)=\frac{\sum_i (r_{ui}-\bar{r}_u) * (r_{vi}-\bar{r}_v)}{\sqrt{\sum_i (r_{ui}-\bar{r}_u)^2}\sqrt{\sum_i (r_{vi}-\bar{r}_v)^2}} \tag{3}$$

其中，u、v、i、r_{ui}、$\sum_i r_{ui}$ 等变量含义同式(2)；$\bar{r}_u$ 与 $\bar{r}_v$ 分别表示用户 u 与 v 对所有物品的评分均值；$\sum_i (r_{ui}-\bar{r}_u)$ 表示用户 u 对物品集合中所有物品的评分集合中的第 i 个物品的评分向量与均值的差值。

1.2.2　基于业务场景改进的协同过滤召回

(1)基于时效性改进的 ItemCF。基于 *item* 的协同过滤算法分析用户对物品的交互记录，找到目标用户曾经喜欢的 *item*，用户对物品的交互记录获取的评分作为 *item* 向量，而不是使用 *item* 本身特征，从物品集合中找出这个喜好 *item* 最相似的 *item*，这个 *item* 将被认为是目标用户的可能喜好，产生推荐。

在新闻推荐场景下，时效性的要求要比普通的物品推荐高得多，为了满足新闻推荐对时效性的需要，在传统基于 item 协同过滤推荐的基础上，设定新闻报道时效的关联规则，将当前时间戳与报道发布的时间戳做差值，当差值超过一定的限度，就不再将该新闻报道纳入被推荐范围内，差值越小，推荐的优先度则越高。为此，本文在相似度计算公式上做出改进，将时效性作为衡量新闻报道相似度的一部分。

在本文中，将 2 个新闻报道的发布时间戳差值的绝对值的倒数作为衡量 2 个物品之间时效性的判断依据，为了防止报道发布时间过近，导致该时效性的值过大影响整个推荐效果，在分母处加 1，防止倒数趋于无穷大。时效性公式如下：

$$h(\boldsymbol{i},\boldsymbol{j})=w*\frac{1}{|i_t-j_t|+1} \tag{4}$$

其中，w 表示时效性的权重；$\boldsymbol{i}$、$\boldsymbol{j}$ 分别表示报道集合中第 i、j 个报道的向量；i_t、j_t 分别表示第 i、j 个报道的发布时间戳。

改进后的基于物品的相似度评判标准可用式(5)来描述：

$$IF(i,j)=sim(i,j)+h(i,j) \tag{5}$$

(2)基于用户活跃度改进的 UserCF。基于 User 的协同过滤运用用户交互构建的评分矩阵，在计算求得用户相似度后，筛选出与目标用户兴趣较一致的用户集合，从这些用户集合中，找出目标用户之前未曾接触过的可能感兴趣的物品，生成推荐序列，推荐给目标用户。

用户集合中有部分活跃用户对新闻报道有较高的关注度，交互记录比较丰富，在评分矩阵中该用户的评分向量较为稠密。活跃用户评分向量的方差较大，覆盖面较为广泛。而普通用户因为个性和偏好的关系，评分矩阵的向量极为稀疏。因此在评价用

户相似时,能够发现大量普通用户都与活跃用户相似,不能很好区分用户是否相似。

为此,本文在用户相似度评判上做出改进,将用户间的活跃度作为比值,活跃度大的作为分母,活跃度较小的作为分子,构建活跃度的判定公式。计算公式如下:

$$q(u,v)=\frac{v_h}{u_h}(u_h \geqslant v_h) \tag{6}$$

其中,u 与 v 分别表示不同用户,u_h 和 v_h 分别表示 2 个用户的活跃度。在新闻推荐的场景下,本文将用户的活跃度设定为用户点击新闻的个数。

将用户活跃度带入用户相似度判定公式中,可得:

$$UF=q(u,v)*sim(u,v) \tag{7}$$

(3)冷启动召回。冷启动召回是在协同过滤召回外的补充,是基于用户特征和物品特征,推断出用户可能感兴趣的物品召回方式。冷启动问题主要分为用户冷启动(即没有用户的交互记录)和新闻冷启动(即新闻没有被任何用户点击)。本文针对冷启动的问题,采用基于热门新闻和地理位置的召回策略,作为额外的新闻召回补充方式。

1.3 算法实现

(1)获取目标用户在新闻平台上的访问日志,包括用户 ID、用户点击的新闻报道 ID,用户点击的时间戳 *Click_timestamp*、该用户点击报道时所在的区域 *Click_region* 等特征。

(2)获取用户曾点击过新闻报道的产生时间戳、以及该新闻报道的字数、新闻报道的类型。

(3)根据用户点击时间戳,存储用户点击新闻的交互记录,构建用户-新闻-时间字典,将其转换为用户评分矩阵。

(4)根据用户评分矩阵,使用改进用户活跃度的 UserCF 查找相似用户,召回新闻报道。

(5)根据特定新闻的用户点击情况,构建新闻报道-用户-时间字典,存储新闻报道的向量,转换成新闻评分矩阵。

(6)根据新闻评分矩阵,使用改进时效性的 ItemCF 来查找相似物品,借此召回新闻报道。

(7)根据用户的点击次数、用户点击报道的区域,进行冷启动召回,召回相关的新闻报道。

(8)汇总多路召回的新闻报道,从中选取评分最高的新闻,作为推荐结果。

综上所述,研究给出算法实现步骤流程如图 2 所示。

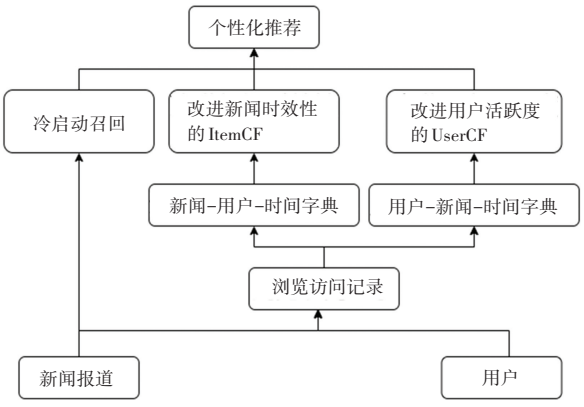

图 2 算法实现

Fig. 2 Flowchart of the algorithm

2 实验及结果分析

2.1 实验设计

为了验证提出个性化新闻推荐的效果,本文数据集选自某新闻平台的后台用户新闻交互访问日志,包含近 30 万用户、36 万篇新闻文章、以及多达 300 多万的交互访问记录。

仿真实验随机从交互数据集中筛选出 2 万用户以及对应的新闻报道和历史交互记录作为数据集,数据集特征包括用户 *id*、用户报道交互时间戳、字数、报道 *id*、报道产生时间戳等特征。构建个性化新闻推荐,针对每个目标用户使用基于用户活跃度的 userCF、基于新闻时效性的 ItemCF 以及冷启动召回等 3 个角度进行召回,每个目标用户将获取个性化召回新闻报道。计算出目标用户对召回新闻的评价,取评价在 Top$X$ 的新闻作为推荐结果,呈现给用户。

2.2 评价指标

为了衡量个性化新闻推荐的推荐效果,本文选取了召回率(*Recall*,R)、准确率(*Precision*,P)以及混合召回和准确率的 F_1 指数作为性能指标。

推荐结果的混淆矩阵见表 1。

表 1 混淆矩阵

Tab. 1 Confusion matrix

	推荐新闻	未推荐新闻
点击	*TP*(推荐给用户且用户点击)	*FN*(未推荐给用户且用户点击)
未点击	*FP*(推荐给用户且用户未点击)	*TN*(未推荐给用户且用户未点击)

选用指标中,*Precision* 表示个性化新闻推荐的准确率,即推荐给用户的新闻列表中,用户点击情况。数学公式定义如下:

$$Precision=\frac{TP}{TP+FP} \tag{8}$$

Recall 表示个性化新闻推荐的召回率，即所有用户点击的新闻列表中，推荐新闻情况。数学计算公式定义如下：

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

由于 *Precision* 和 *Recall* 两个指标在某些情况下是相互冲突的、准确率高、召回率就会降低，召回率高、准确率会降低。为了平衡 2 个指数的冲突，更好地衡量个性化推荐的推荐效果，引入了 F_1 指标。数学公式定义如下：

$$F_1 = \frac{2PrecisionRecall}{Precision + Recall} \tag{10}$$

2.3　实验结果与分析

本次实验从 20 000 个用户的数据集中随机选取某用户，分别以 itemCF 和 UserCF 以及个性化推荐这 3 种方式针对该用户从其最相邻用户集合中选出 5、10、15、20、25、30、35、40 个邻居，基于最近邻居产生目标用户的个性化推荐。

表 2 和表 3 是个性化推荐算法与基于 User 和 Item 协同过滤推荐算法的针对目标用户产生推荐的实验效果对比。F_1 – *score* 的实验对比结果见图 3。从以上结果中可以看出，在协同过滤的基础上加入用户活跃度和新闻时效性的考量，使用冷启动补充兴趣点，在新闻场景下的推荐效果明显优于协同过滤。

表 2　准确率实验对比

Tab. 2　Comparison of accuracy experiments

P	Item	User	CUI+Cold
5	0.244 88	0.270 77	0.281 63
10	0.249 64	0.264 28	0.297 40
15	0.254 10	0.268 96	0.290 89
20	0.246 26	0.263 85	0.295 80
25	0.244 84	0.275 51	0.290 34
30	0.244 66	0.267 26	0.293 26
35	0.242 35	0.265 84	0.294 96
40	0.248 56	0.263 16	0.290 35

表 3　召回率实验对比

Tab. 3　Comparison of recall experiments

R	Item	User	CUI+Cold
5	0.064 57	0.072 00	0.073 80
10	0.065 07	0.070 44	0.074 34
15	0.065 34	0.072 71	0.074 31
20	0.064 90	0.072 84	0.076 10
25	0.066 34	0.074 23	0.075 63
30	0.065 85	0.071 01	0.076 30
35	0.065 52	0.069 99	0.077 21
40	0.066 21	0.068 70	0.075 86

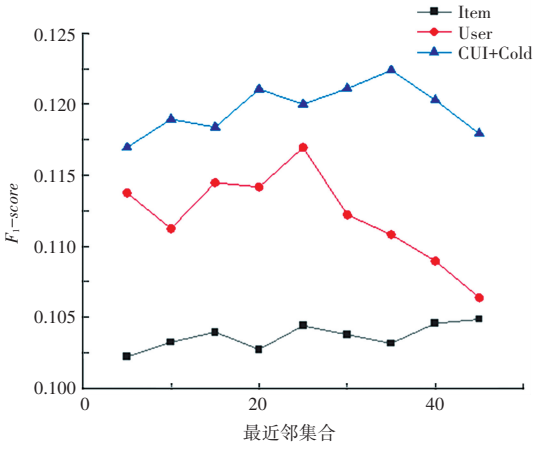

图 3　F_1–*score* 实验对比

Fig. 3　F_1–*score* experimental comparison

3　结束语

本文基于用户活跃度和新闻时效性提出了一种个性化的新闻推荐算法，考虑到新闻推荐场景下，用户对新闻时效性的需求以及不同活跃度用户对推荐的影响，修改了新闻报道和不同活跃度用户的相似度判定方式，基于协同过滤，从用户、新闻报道和冷启动召回的角度召回用户可能感兴趣的新闻报道。实验证明，与传统的协同过滤推荐相比，本文的个性化新闻推荐更为准确。冷启动召回的方式缓解了冷启动，而改进后的召回排序获取推荐提高了推荐精度。下一步可以在新闻推荐中融合、更多模型算法更多覆盖用户的兴趣点完成召回和排序，提高推荐效果。

参考文献

[1] 万梅，曹琳. 基于神经网络嵌入和动态社交的新闻推荐算法[J]. 计算机应用与软件，2021，38(07)：258-264+331.
[2] 刘羽茜，刘玉奇，张宗霖，等. 注入注意力机制的深度特征融合新闻推荐模型[J]. 计算机应用，2022，42(02)：426-432.
[3] 张丹，周从华. 基于改进的隐马尔可夫模型的新闻推荐算法[J]. 计算机与数字工程，2020，48(10)：2332-2337.
[4] 袁仁进，陈刚，李锋. 面向新闻推荐的用户兴趣模型构建与更新[J]. 计算机应用研究，2019，36(12)：3593-3596.
[5] 陶天一，王清钦，付聿炜，等. 基于知识图谱的金融新闻个性化推荐算法[J]. 计算机工程，2021，47(06)：98-103，114.
[6] 王利娥，李小聪，刘红翼. 融合知识图谱和差分隐私的新闻推荐方法[J]. 计算机应用，2022，42(05)：1339-1346.
[7] 朱文跃，刘炜，刘宗田. 基于事件本体的新闻个性化推荐[J]. 计算机工程，2019，45(06)：267-272，279.
[8] 朱郁筱，吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报，2012，41(02)：163-175.