基于 TF-IDF 算法在协同过滤推荐算法的应用研究

邵茂仁 杨晓飞 林丽婷

摘　要：个性化推荐系统现今愈加被重视和使用，但其存在的稀疏性问题也愈加凸显。为改善稀疏性问题现大部分学者和研究员对此提出了许多基于矩阵分解和矩阵填充等其他角度的解决方案，鲜有观点提出从信息检索角度重新看待此问题。文本针对稀疏性问题，提出了基于协同过滤核心思想，从信息检索角度利用该领域技术将“预测推荐问题”转化为“信息检索分类”问题，并基于 MovieLens 数据集设计了对比实验。实验表明，本文提出的解决方案在 MRR 指标下的表现高于传统协同过滤算法在该指标中的表现。

关键词：协同过滤；稀疏性；TF-IDF 算法

Application of TF-IDF algorithm in Collaborative Filtering Recommendation Algorithm

Shao Maoren　Yang Xiaofei　Lin Liting

Abstract: Personalized recommendation system is more and more valued and used nowadays, but its sparse problem is becoming more and more prominent. In order to improve the sparsity problem, most scholars and researchers have put forward many solutions based on matrix decomposition and matrix filling. Few viewpoints have proposed to re-view this problem from the perspective of information retrieval. Aiming at the sparsity problem, this paper puts forward the core idea of collaborative filtering. From the perspective of information retrieval, the technology of this field is used to transform "predictive recommendation problem" into "information retrieval classification" problem, and a comparative experiment is designed based on MovieLens data set. Experiments show that the performance of the proposed

solution under MRR index is higher than that of traditional collaborative filtering algorithm in this index.
Key Words: Collaborative filtering; sparsity; TF-IDF algorithm

0 引言

伴随着信息时代、数据时代的到来、海量信息的产生，这一方面给用户提供了更加丰富多元的选择，另一方面也增加了用户的选择难度。而智能化、个性化的推荐服务因其能为用户达到主动推送个性化物品、一定程度上提高用户体验的目的，已逐渐成为各领域产业尤其是电商产业的极大需求[1]。传统的个性化推荐系统以协同过滤算法为代表的推荐算法已经应用在了实际的业务中并取得了一定的效果。但随着数据的剧增，传统协同过滤算法中稀疏性问题也日渐凸显加剧。

稀疏性问题主要是由于用户数量与物品数量增加但用户对物品的评价数量却没有以同等的速率增加而产生的，同时稀疏性问题也是导致推荐质量下降的主要原因。而现阶段该问题的解决方案主要分为矩阵分解、矩阵填充和基于聚类、标签划分等其他解决方案[2-4]。

本文针对传统协同过滤算法中的核心推荐原理思想和稀疏性问题进行了相关研究，提出了：基于协同过滤算法核心思想，从信息检索的角度将“推荐预测”问题转化为“信息检索分类”问题的解决方案。利用 TF-IDF 算法相关原理提取出目标用户的相似用户群，进而产生相应的推荐列表。通过对比实验表明，本文提出的方案在解决了稀疏性问题的前提下算法产生的推荐列表在 MRR 指标值中的表现高于传统协同过滤算法在该指标中的表现。

1 相关技术介绍

1.1 传统协同过滤推荐算法介绍

协同过滤算法作为推荐技术中使用最早且最知名的推荐算法，其在学术界得到了较深入的研究以及在实际产业中的到较多的应用。其中，协同过滤可分为基于用户的协同过滤（UserCF)和基于物品的协同过滤（ItemCF）。其核心思想遵循“趣味相投”的原理，即找到与目标用户相似的用户群，然后将该用户

群中的相似用户喜欢而目标用户未评价过的物品推荐给目标用户。

传统的协同过滤算法具体实现主要分为四个步骤：

（1）过收集所有用户的历史评分数据，构建用户-评分矩阵。

（2）根据构建的用户-评分矩阵进行用户间的相似度计算，构建用户相似度矩阵。

（3）根据用户相似度矩阵找出与目标用户相似的用户群。

（4）将相似用户群中目标用户未评价过的物品形成推荐列表推荐给目标用户[5]。

由于用户评分较少，传统协同过滤算法中的稀疏性问题在步骤一构建用户-评分矩阵时便会体现出来。

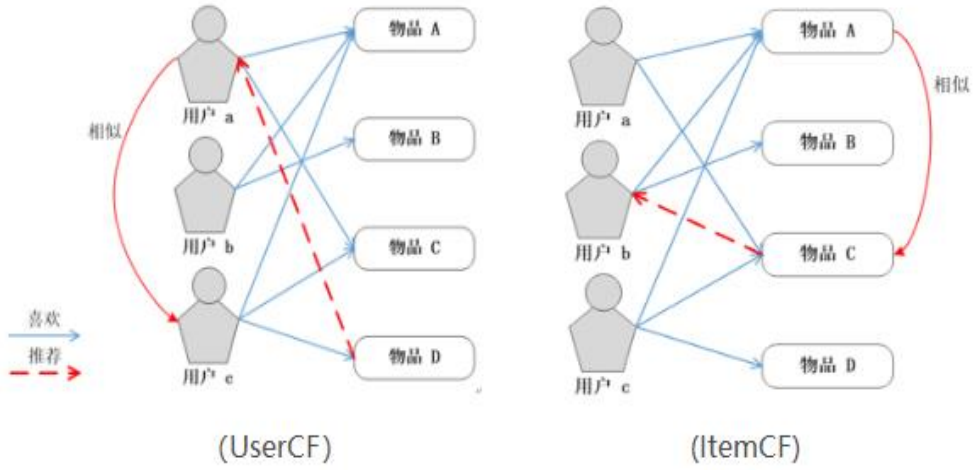

图 1 协同的过滤算法基本原理

1.2 TF-IDF 算法介绍

TF-IDF（term frequency－inverse document frequency）算法是一种基于词频 TF(Term Frequency)与逆文档频率指数 IDF(Inverse Document Frequency)的统计方法[6],一种用于信息检索与数据挖掘的常用加权技术。其主要的核心思想是：如果某个特征词 w 在文档 d 中出现的频率很高同时出现在少量的文档中，那么可认为该特征词对于该篇文章具有很好的区分能力[7]。其中 TF-IDF 的计算公式如下：

$$TF-IDF(\mathrm{w}_i)=tf(w_i)\times idf(w_i)$$
$$=tf_j(w_i)\times\log(N/df(w_i)) \quad (1)$$

其中 $\mathrm{tf_j}(w_i)$ 的含义为当前关键词 w_i 在文本 j 中出现的频率，N 则表示总文章数，$\mathrm{df}(w_i)$ 则表示的是当前关键词 w_i 出现在了多少篇文章中[6]。

2 算法设计介绍

2.1 算法内容描述

针对传统协同过滤算法的核心思想研究后发现，其关键内容在于找到与目标用户相似的相似用户群，进而为目标用户进行推荐。而“推荐”这一动作本身，从信息检索的角度看待，与“搜索”这一动作本身具有相似性。两者表面上看似不同：推荐是由系统自动完成，由系统主动将个性化的内容呈现给用户；搜索，是由用户主动进行一定的输入然后系统再将相似的内容主动呈现给用户。但是，从搜索的角度，如果将系统本身的一部分抽象为“用户”让系统进行主动输入然后再经由推荐系统进行推荐内容的输出，则可将“推荐”与“搜索”这两个动作近似相等，那么，在信息检索中所运用的技术逻辑上也可运用于推荐系统中。本文提出的解决方案核心算法思想便是由此而推理形成。

与传统协同过滤算法相比，本文的核心算法并没有构建用户-评分矩阵，同时也没有像传统协同过滤算法中使用余弦相似度、基于皮尔逊相关系数等方式进行用户相似度的计算。而是将“寻找与目标用户相似的用户群”这一问题从信息检索的角度将其从一个“推荐预测”问题转化为一个“信息检索分类”问题，即转化为：计算目标用户的评分物品同时也“属于”其他用户的评分物品的可能性是多少，将可能性大的用户定义为目标用户的相似用户群，进而产生最后的推荐列表给目标用户。

2.2 算法伪代码描述

表 1 算法伪代码符号说明

符号	说明
user_id	用户 id
sim_user_n	相似用户个数
top_k	推荐列表长度
top_k_rec_list	最终推荐列表结果
get_user_items_list()	获得目标用户的评分列表
user_items_list	目标用户的评分列表
get_sim_users_list()	获取相似用户列表
sim_users_list	用户相似列表
get_top_n_sim_user	获取前 n 个相似用户
get_top_k_rec_list()	获取长度为 k 的最终推荐列表
tfidf_score_dic	存放 item 在某个 user 中的 tfidf 值字典

算法 1 核心推荐算法 get_rec()

输入: user_id, sim_user_n, top_k

输出: top_k_rec_list

(1) user_items_list = get_user_items_list(user_id)

(2) sim_users_list = get_sim_users_list(user_items_list, user_id)

(3) get_top_n_sim_user = sim_users_list[:sim_user_n]

(4) top_k_rec_list = get_top_k_rec_list(user_id,get_top_n_sim_user, top_k)

算法 2 获取相似用户列表 get_sim_users_list()

输入: user_items_list, user_id

输出: sim_users_list

(1) tfidf_score_dic = {}

(2) for item in user_items_list:

(3) 获取拥有该 item 的 item_user_id_list

(4) for item_user_id in item_user_id_list:

(5) 分别计算每个 item 在拥有该 item 的 user 中的 tfidf 值 v

(6) tfidf_score_dic[item_user_id] = v

(7) end

(8) end

(9) 排序 tfidf_score_dic

(10) return sim_users_list

3 实验流程

为客观评价本文提出的解决方案与传统协同过滤算法的优劣，本文基于控制变量法设计了对比实验。实验流程主要分为：数据预处理、模型训练、结果预测三个部分。

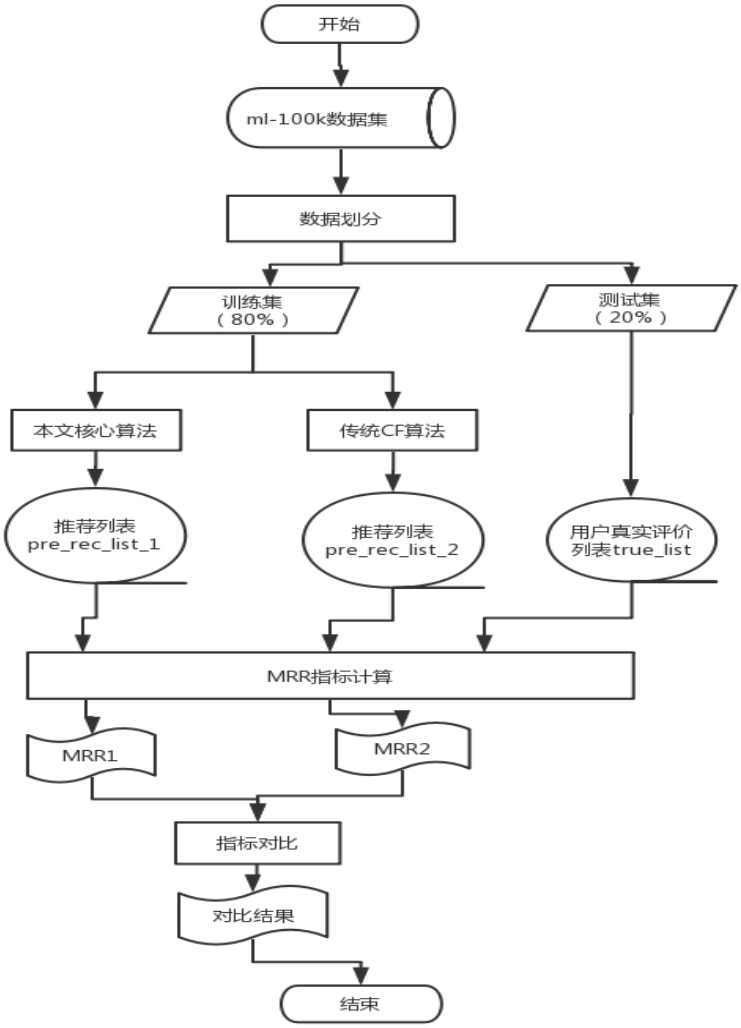

图 2 对比实验流程图

3.1 数据预处理

本实验所采用的是公开的 MovieLens 数据集 ml-100k，以 8:2 的比例划分成为训练集和测试集两个部分。以测试集中用户的评分项列表作为仿真测试结果，以训练集中的用户行为数据作为训练数据进行模型的训练。

3.2 模型训练

在注入用户行为数据后，模型主要计算每个用户评分项目中各个项目在整个用户行为数据中的 TF-IDF 值，并与对应的用户 ID 进行映射，将计算结果存入数据库中以待下一步查询计算使用。

3.3 结果预测

本实验对于结果的预测，主要分为以下几个步骤：

Step1：获取测试集中的目标用户 id 列表 true_user_id_list。

Step2：输入目标用户 ID，获取目标用户的推荐列表 pre_rec_list。

Step3：计算 pre_rec_list 与 true_user_id_list 的匹配排名倒数值。

Step4：重复 Step2、Step3 直到给测试集中的每个目标用户完成推荐。

Step5：取 Step3 中匹配排名倒数值计算模型的 MRR 指标值。

4 实验结果及分析

4.1 采用的数据集介绍

本文设计的对比实验采用的 ml-100k 数据集源于 MovieLens 网站（movielens.umn.edu），是一份开源数据集。该数据集包括了 1682 部电影的 943 名用户获得 100,000 点评分（1-5 分）等[8]。该数据集已经以 8:2 的比例划分成为了训练集和测试集。该数据集也常被用于个性化推荐系统的研究当中。

4.2 实验环境

表 2 实验软硬件配置信息表

软/硬件	配置/说明
处理器	英特尔 Core i5-4210M @2.60GHz 双核
运行内存	8 GB（金士顿 DDR3L 1600MHz）
操作系统	Windows10 专业版 64 位
编程语言	Python3.6.2
数据库	Sqlite3

4.3 评价指标

自推荐算法研究以来，对于推荐系统的推荐效果的评价指标可以从推荐的准确度、推荐的多样性、新颖性、基于信息检索系统等角度进行评价[9]。其中较为常用的评价指标有平均绝对误差（MAE）、平均平方误差（MSE）、均方根

误差（RMSE）、ROC 曲线面积（AUC）等。这些评估指标多是基于评估推荐系统对推荐物品所估计的偏好值在多大程度上与实际偏好值的匹配。

本文提出的核心推荐算法并未对偏好值进行评估，最终结果只输出了一个排序后的推荐列表。为在对比实验中控制其他变量不变，使得对比结果更具有说服力，综合以上因素本文在评价指标方面采用了平均倒数排名(Mean Recipprocal Rand，MRR)进行推荐系统的评价。其中，MRR 定义如下：

$$MRR = \frac{\sum_{q=1}^{Q} 1/rank_i}{Q} \tag{2}$$

其中 Q 为系统为物品推荐的个数，$rank_i$ 为目标用户仿真项目列表中的物品在推荐列表中的排名[10]。从 MRR 的定义中可看出，如果 MRR 的值越大，那么推荐系统的性能就越好。

4.4 实验结果

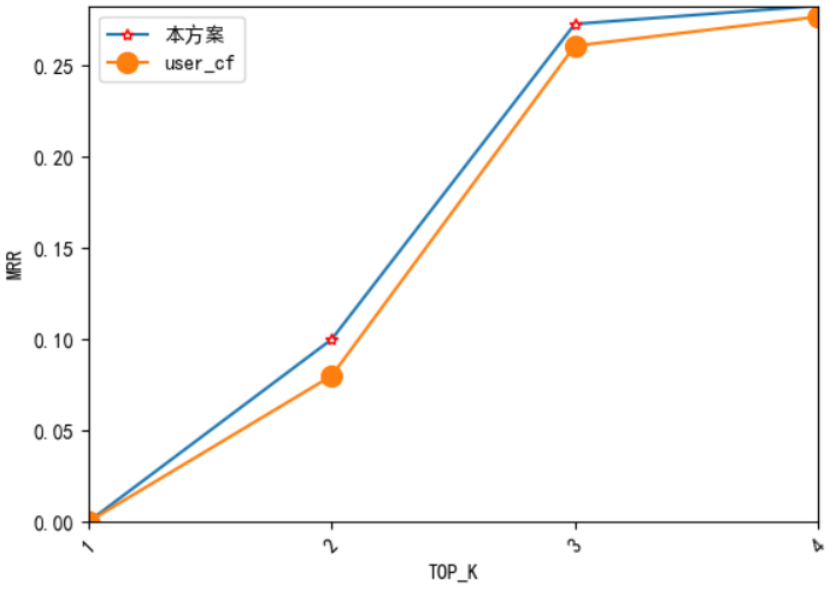

图 3：不同推荐列表长度下的 MRR 值

本文设计了四组对比实验，将推荐列表的长度分别设计成了 1、2、3、4 个推荐。通过实验可看出本文提出的解决方案的 MRR 值略高于传统 user_cf 算法。

5 总结与展望

本文首先对个性化推荐算法进行了简要概述以及分析了传统推荐算法中稀疏性问题的成因并对现阶段为此而提出的改善方案进行了简述,然后提出了本文的解决方案：基于协同过滤算法核心思想，从信息检索的角度将“推荐预测”问题转化为“信息检索分类”问题。接着介绍了本文提出的解决方案所运用到的核心技术并对核心算法进行伪代码介绍,最后介绍了本文的对比实验的设计及实验过程并展示了对比实验的结果。

通过对比实验表明：本方案在规避了稀疏性问题的前提下，在 MRR 指标值下的表现高于传统协同过滤算法下的 MRR 指标值。本文最重要的创新点和突出贡献主要有以下几点：

1. 舍弃传统推荐算法中相似用户群的计算方式，充分利用了已有数据集中的数据信息，运用信息检索方面的技术获取到了目标用户的相似用户群，最终产生推荐列表给目标用户。

2. 在“推荐预测”问题上通过抽象出系统本身的一部分作为一个特殊的“用户”主动输入用户项目信息后，将该问题转化为“信息检索分类”问题，从而进一步利用信息检索方面的技术进行问题的解答,为日后更深入的研究提供了一个方向。

然而本文在研究和实验过程中也发现了该方案存在的一些不足其中包括：在处理较大规模数据时运行时间长、所运用的信息检索技术效果非最佳等问题。对此，在日后的研究中可从优化算法结构、将系统搭建在 spark 平台上运行、更加充分利用数据集中的信息尝试构建“特征-标签矩阵”使用 ML、DL 技术等角度和方法进行更深层次的研究。

参考文献：

[1]段国仑,谢钧,郭蕾蕾,王晓莹.Web 文档分类中 TFIDF 特征选择算法的改进[J/OL].计算机技术与发展,2019(05):1-4[2019-03-11].http://kns.cnki.net/kcms/detail/61.1450.TP.20181221.1554.052.html.

[2]赵宇峰,李新卫.基于歌曲标签聚类的协同过滤推荐算法的研究[J]
.计算机应用与软件,2018,35(06):259-262.
[3]孙华艳,李业丽,宇云飞,韩旭.协同过滤推荐算法的改进与研究[J]
.计算机技术与发展,2018,28(10):44-48.
[4]黄贤英,龙姝言,谢晋.结合用户兴趣度聚类的协同过滤推荐算法[J/OL]
.计算机应用研究,2019(09):1-7[2019-03-11]
.http://kns.cnki.net/kcms/detail/51.1196.TP.20180524.2109.032.html.
[5]王冲. 基于SVD和用户聚类的协同过滤算法研究[D]
.青岛理工大学,2018.
[6]黄瑛. 基于用户聚类和偏好的推荐算法研究[D].安徽理工大学,2017.
[7]黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J].计算机学报,2011,34(05):856-864.
[8]F.Maxwell Harper and Joseph A.Konstan. 2015. The MovieLens Datasets:History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.DOI=http://dx.doi.org/10.1145/2827872
[9]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.
[10]刘攀,陈敏刚.个性化推荐系统评估[J]
.南昌大学学报(理科版),2016,40(02):143-150.