流行度划分结合平均偏好权重的协同过滤个性化推荐算法

何佶星 陈汶滨 牟斌皓
(西南石油大学计算机科学学院 成都 610500)

摘 要 提出了一种考虑平均偏好权重的协同过滤个性化推荐算法。该算法分为邻域计算、数据集划分、偏好预测 3 个阶段。在邻域计算阶段,采用基于欧氏距离的 KNN 来确定邻域;同时对数据集按照其本身特点设定的流行度阈值进行划分;在预测评分时,对已有的邻域按照流行度选取部分项目,基于项目集的偏好相似度求解用户的平均偏好权重,据此对用户进行先后两次预测,再求平均结果。在 Movielens 100K 数据集上将所提算法与典型的余弦推荐算法、person 推荐算法、基于项目偏好的协调过滤算法和用户属性加权活跃近邻的协同过滤算法进行比较实验,结果表明新算法在 MAE 上表现的更优秀。

关键词 协同过滤,KNN,个性化推荐算法,流行度划分,平均偏好权重,邻域计算
中图法分类号 TP391 **文献标识码** A

Coordination Filtering Personalized Recommendation Algorithm Considering Average Preference Weight and Popularity Division

HE Ji-xing CHEN Wen-bin MOU Bin-hao
(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract This paper presented a new recommendation algorithm which takes into account the average preference weight. The algorithm is divided into three stages: neighborhood computing, data set partitioning and preference prediction. In the neighborhood calculation, the KNN based on the Euclidean distance is used to determine the neighborhood. At the same time, the data set is divided into the data set and the non-popular data set according to the popularity threshold of the data set itself. When the score is predicted, the existing neighborhood selects part of the project according to the popularity degree, and predicts the user's average preference weight based on the preference similarity of the item set. The results show that on the Movielens 100K data set, the new algorithm is superior to the typical cosine recommendation algorithm, the person recommendation algorithm, the collaborative filtering algorithm based on the project preference coordination filtering algorithm and the user attribute weighted active neighbor existing algorithms in MAE.

Keywords Coordination filtering, KNN, Personalized recommended algorithm, Popularity division, Average preference weight, Neighborhood calculation

在电子商务以及社交网络不断发展的今天,用户在通过搜索引擎检索所需信息的同时,也面临数据冗余、信息超载检索困难的问题,庞大的信息库使得用户很难迅速找到需要的信息或者服务。这时就要借助推荐功能来为用户进行个性化推荐。传统的用户推荐算法的推荐准确度较低一直是一个亟待解决的问题。为不同的用户精准地推送信息,提高数据的利用效率是推荐算法的核心目标。

目前,个性化推荐技术中,基于协同过滤(Collaborative Filtering,CF)[1-3]的推荐算法主要分为基于用户的推荐和基于项目的推荐两种。基于用户的协同过滤推荐技术是基于已有用户的评分来预测目标用户的项目评分,从而得到 top-N 的项目推荐[4-8]。基于的项目的推荐技术通过对项目信息的特征进行解析,计算其与目标用户偏好的符合程度,从而推荐项目。文献[9]针对传统协同过滤算法不能通过深度挖掘用户关系对新项目进行推荐的问题,提出了通过计算用户评分的活跃度和项目评价结果对平价均值的偏差来计算用户之间相似度的协同过滤算法。但是,用户评分活跃度容易受到矩阵稀疏度的影响。文献[10]对于现有的 KNN 协同过滤算法过于依赖评分相似度的问题,提出了一种在已有领域中寻找最活跃领域来确定预测评分的协同过滤算法,然而其结果缩小了已有领域,对精确度的提升较小,活跃用户在预测评分时缺乏主次之分。

基于上述分析,本文提出一种结合流行度划分的新型偏好权重的协同过滤个性化推荐方案。首先,使用基于用户的 K 最近邻(K-Nearest Neighbor,KNN)[11-13]推荐算法获取基于用户近邻用户,获得邻居用户的对应的用户-项目-评分矩阵表;然后,通过流行度划分改进数据集的输入方式;最后,使用一种新的基于项目评分正确度的计算近邻用户偏好匹配权重的方式对目标用户的目标项目进行预测评分。两种方法分步结合可以在用户评分数据稀疏的情况下提高预测的准确

何佶星(1991—),女,硕士生,CCF 会员,主要研究方向为推荐算法、机器学习等,E-mail:jqay1234@126.com(通信作者);**陈汶滨**(1965—),男,教授,硕士生导师,主要研究方向为油田信息化、数据库技术与应用、计算机模拟与仿真;**牟斌皓**(1992—),男,硕士生,CCF 会员,主要研究方向为数据挖掘和机器学习。

度。同时,本文提出了用户偏好权重和数据集流行度,使两者能够更好地结合,优化了基于用户的协同过滤的结果。实验结果表明,该方案能够更加准确地预测出项目评分,提高个性化推荐技术的精准度。本文基于 MovieLens 公开数据集,通过向用户推荐电影,对用户的评分进行预测。

1 问题描述

1.1 流行度数据集模型

在互联网中,用户对物品的行为有很多种,评分是应用最广泛的方式,常用的三元关系组为 U(用户),I(项目),R(评分)。用户-项目-评分矩阵如表 1 所列。

表 1 用户-项目-评分矩阵

	项目 1	…	项目 j
用户 1	R_{11}	…	R_{1j}
…	…	…	…
用户 i	R_{i1}	…	R_{ij}

定义 1 基于项目流行度的评分数据集是一个五元组:

$$S=(U,V,R,r,p) \tag{1}$$

其中,$U=\{u_1,u_2,\cdots,u_n\}$是用户的有限集合,$V=\{v_1,v_2,\cdots,v_m\}$是项目的有限集合,$R=\{1,2,3,4,5\}$是评分的有限集合,$r:U\times V\rightarrow R$ 是用户 $u\in U$ 对项目 $v\in V$ 的评分函数,记为 $r_{u,v}$;$p:V\rightarrow N^+\cup\{0\}$是项目 $v\in V$ 的流行度函数。

值得注意的是,当 $r_{u,v}=0$ 时,本文认定用户 u 与项目 v 之间没有产生联系,如不曾观看该电影或不曾使用该商品。

定义 2 给定项目流行度的评分数据集 $S=(U,V,R,r,p)$,项目 $v\in V$ 的流行度为:

$$p_v=|\{r_{u,v}|r_{u,v}>0,u\in U\}| \tag{2}$$

表 2 是一个项目流行度的评分数据集实例,共列举了 5 名用户对 3 部电影的评分状态以及每部电影的流行度。刘一对电影《泰坦尼克号》的评分为 4 分,且没有观看其余 2 部电影。电影《金刚》的流行度为 3,因为共有 3 名用户观看;而电影《泰坦尼克号》的流行度为 5,因为它被当前所有所用户观看过。

表 2 流行度评分数据模型

(a)评分数据集实例

用户	泰坦尼克号	金刚	美女与野兽
刘一	4	0	0
陈二	5	4	0
张三	3	0	1
李四	4	1	2
王五	3	3	0

(b)电影流行度

电影	泰坦尼克号	金刚	美女与野兽
流行度	5	3	2

为了对海量的项目进行区分,进而获得高流行度的项目集(热门项目集)和低流行度的项目集(冷门项目集),本文引入流行度阈值的定义。

定义 3 给定项目流行度的评分数据集 $S=(U,V,R,r,p)$和项目流行度阈值(popularity threshold)

$$pt=\frac{\mathrm{Max}(p_v)}{2} \tag{3}$$

进而可得高流行度项目集为:

$$V_H=\{v\in V|p_v\geqslant pt\} \tag{4}$$

低流行度项目集为:

$$V_L=\{v\in V|pv<pt\} \tag{5}$$

例如,由定义 3 可知表 2 的流行度阈值 $pt=\frac{Max(5,3,2)}{2}=2.5$,因此,高流行度项目集 $V_H=${泰坦尼克号,金刚},低流行度项目集 $V_L=${美女与野兽}。

1.2 相似度模型

KNN(近邻)算法最初由 Hart 和 Cover 于 1968 年提出,此后经过逐步的研究应用,实践结果表明 KNN 相较于其他算法更加适用于类域相交或重叠较多的待处理样本集。本算法使用了 k 个邻居的概念,把目标用户和 k 个邻居归类为同一个类,然后利用已知的邻居用户来预测目标用户对目标电影的评分,从而判定用户对该电影的偏好程度。KNN 实现简单并且其分类准确度较高。

主要的推荐方法有两大类:基于相似度的推荐和基于用户距离的推荐技术。对于类似 Netflix 豆瓣电影的电影推荐,已有推荐系统大多会在观看一部电影后对电影进行打分,分值为 1~5 的整数,从而表现出用户对电影的偏好程度。根据用户对电影的评分来选取预测结果分值偏高的电影推送给相应的用户。

定义 4 用户集中,用户 u 对商品 a 的预测评分为:

$$R_{ua}'=\overline{R_u}+k\sum_v S_{uv}(R_{va}-\overline{R_v}) \tag{6}$$

其中,U 代表用户数据集,$(u,v)\in U$;R_{ua} 代表用户 u 对项目 a 的评分。$\overline{R_u}$和$\overline{R_v}$表示用户 u 和用户 v 的平均评分;$k=(\sum_v S_{uv})^{-1}$为矫正因子,其中 v 代表所有对商品 a 进行评分的用户,S_{uv}代表用户 u 与其他用户之间的相似度。

1.2.1 基于用户相似度的推荐

基于用户的协同过滤算法一般计算其余用户与测试用户的相似度 $sim(u,u_1),sim(u,u_2),sim(u,u_3),\cdots,sim(u,u_n)$的值并且排序,相似度计算方法一般有以下几种。

定义 5 用户 u 和用户 v 之间的 cosine 距离的用户相似度为:

$$sim(u,v)=\cos(u,v)=\frac{\sum_{i=1}^{n}R_{u,i}\times R_{v,i}}{\sqrt{(\sum_{i=1}^{n}R_{u,i})^2\times(\sum_{i=1}^{n}R_{v,i})^2}} \tag{7}$$

相较于余弦距离,person 系数的用户相似度还考虑了用户的评分与平均值之间的分差。

定义 6 用户 v 和用户 u 之间的 person 相似度为:

$$sim(v,u)=\frac{\sum_{i=1}^{n}(R_{v,i}-\overline{R_v})(R_{u,i}-\overline{R_u})}{\sqrt{\sum_{i=1}^{n}(R_{v,i}-\overline{R_v})^2\times\sum_{i=1}^{n}(R_{u,i}-\overline{R_u})^2}} \tag{8}$$

其中,$R_{v,i}$为用户 v 对项目 i 的评分,$\overline{R_v}$为用户 v 对 k 个项目的评分的平均值。

1.3 评分预测

已有协同过滤推荐算法一般在寻找邻域的计算方法上进行研究,进而对用户评分进行预测。本文提出的已有邻域用户基于邻域用户的项目集的匹配度来对用户分别加权,从而对邻域用户的进一步处理分两次计算,最终得到预测评分。

1.3.1 项目正确度

本文提出了其余用户对目标用户评分的正确度系数,由于每个用户的每一项评分在测试用户中对应项目的正确度与

项目间的错误度互补，取项目错误因子为项目之间的偏差在分值上线中占的比例，在协同过滤推荐系统中，如表 1 中 $i\times j$ 用户项目评分表。

定义 7　假设存在用户 u 和用户 v，则用户 v 对用户 u 关于项目 j 的正确度为：

$$Right(u,v,j)=1-\frac{|R_{v,j}-R_{u,j}|}{full} \tag{9}$$

其中，$R_{v,j}$ 与 $R_{u,jn}$ 同时存在且不为 0。

1.3.2　偏好权重定义

将基于协同过滤相似度得出的已有邻居按照其重要程度分别加权，削弱因邻域用户质量不统一而对预测结果的影响。基于用户偏好权重的协同过滤算法依据来源于邻居用户对项目集的评分正确度的加权平均值、邻域评分矩阵和测试用户共同项目的评分，在一定程度上反映了邻居用户对于项目集的共同偏好程度，两者之间的偏差值的比重可以代表两者对共同项目喜好度的偏差权重。其在评分区域的补集记为两者之间的偏好正确度，项目集合的正确度的平均值表示每个领域用户的平均偏好权重。每个用户的权重都是独立于其他领域用户即又与测试用户相关联的，根据每个用户的权重计算测试用户的预测评分更加具有合理性。

定义 8　偏好权重是用户之间的项目评分的拟合程度。计算偏好权重的步骤如下。

第 1 步：通过计算 euclidean Distance 求得邻域用户。用户 u 与用户 v 间的 euclidean Distance 的表达式如下：

$$D_{u,v}=\sqrt{\sum_{i=1}^{n}(R_{u,i}-R_{v,i})^2}/k \tag{10}$$

第 2 步：计算邻域用户 v 对用户 u 的项目正确度集，取平均值，从而求出每个近邻用户的算术平均权重。其表达式如下：

$$(Weight)_k=\frac{\sum_{i=1}^{n}Right_i}{M_{count}} \tag{11}$$

其中，n 为除了目标电影以外的其余用户与邻居同时观看过的电影数目。

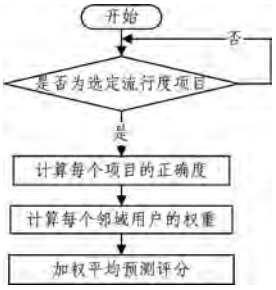

图 1　偏好权重计算流程图

1.4　评价指标

本文采用平均绝对误差作为算法的评价指标。相较于平均误差而言，其离差被绝对值化，没有正负相抵消的情况，可更好地反映预测值误差的实际情况。

定义 9　平均绝对误差（Mean Absolute Error，MAE），是所有单个观测值与算术平均值的偏差的绝对值的平均。

$$MAE=\frac{1}{E^p}\sum_{(u,a)\in E^p}|R_{ua}-R_{ua}'| \tag{12}$$

其中，R_{ua} 表示用户 u 对商品 a 的真实评分，R_{ua}' 表示用户 u 对商品 a 的预测评分，E^p 表示测试集。

2　算法描述

2.1　划分数据集

输入：待分类的用户-项目-评分混合数据集 $S=\{Item_1,Item_2,\cdots,Item_j\}$（j 项目个数）

输出：基于流行度的电影项目的类

步骤 1　计算所有数据集的流行度阈值；

步骤 2　划分数据集。

预测评分时的输入数据集分别为流行数据集（popMatrix）和不流行数据集（UnpopMatrix），用户数目不变。即分别只考虑流行度高的项目在相同领域对预测项的偏好加权和流行度低的项目的偏好加权，每次的预测结果有两个，将其平均求和可得总体结果。本文所提算法在没有改变输入的基础上在计算中将输入数据分割，然后分别进行计算。

2.2　算法流程描述

为了解决用户评分矩阵稀疏导致的预测准确度较低的问题，本文将基于内容的协同过滤技术与基于用户的协同过滤推荐的新算法（Average Preference Weight algorithm，APW）相结合，利用基于用户 KNN 的推荐算法获得邻域。再结合基于流行度的方法来对数据集进行划分，获得相关数据集后结合新提出的 APW 算法对项目做出最终评分，从而对用户进行推荐。所提算法的流程如图 2 所示。

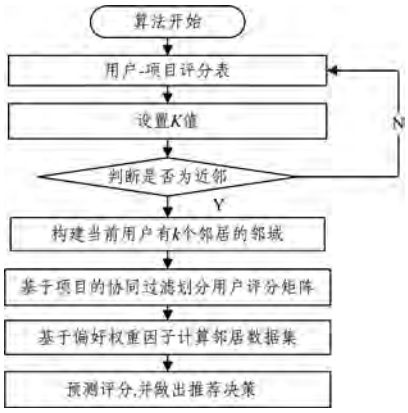

图 2　AWP 算法流程图

由于基于用户的协同过滤技术采用的算法对已经找出的近邻用户数据集没有主次之分，在预测计算时会产生一定的误差，因此提出 APW 算法以修正这种偏差，从而提高推荐精度。

输入：相似用户的电影评分矩阵 $S=\{R_{11},R_{i2},\cdots,R_{kj}\}$，j 为项目个数，k 为邻居个数，R 为用户对电影的评分

输出：获得的电影推荐

目的：提高数据集预测精准度。

步骤 1　读入已找出的相似用户的评分矩阵。

步骤 2　分别计算除去目标用户以外的其余用户对除去目标电影以外的其余电影的评分偏好权重。

步骤 3　引用偏好权重系数进行加权平均，对评分值进行预测。

$$\text{Predict}=\frac{\sum_{i=1}^{k}\text{Weight}_k\times R_{ki}}{\sum_{i=1}^{k}\text{Weight}_k} \tag{13}$$

步骤 4　对预测评分排序，获得推荐。

3　实验设置

构建实验环境，对不同的推荐算法进行评估。实验配置

为 Intel 酷睿 i5 处理器、4 GB RAM、WIN8 系统。通过基于 eclipse平台的 JAVA 编程语言实现推荐算法。采用 GroupLens 项目组提供的 MovieLens 电影数据库，其中包含用户对影片评分，即 943 个用户对 1682 部电影的 rating 数据，rating数据中包括用户编号 userId 和电影编号 movieId，评分范围为 1～5，0 分代表用户尚未对电影进行评分。数据集的稀疏性为 1－100000/943×1682＝0.93。得到用户的相似度以后，本实验采取 leave one out 方法，每次采用一个用户评分作为测试集，其余数据作为训练集。

实验中，将用户对电影的预测评分与实际评分的平均绝对误差（Mean Absolute Error，MAE）[14] 作为评价指标。MAE 值越小，说明算法的推荐准确性越高。

3.1 性能测试

在数据集上执行本文提出的推荐算法，以验证邻域用户数量对算法推荐性能的影响。设置邻域用户数量 k 的范围为 1～80。同时将数据集按照流行度保存为流行数据集 popMatrix 和不流行的数据集 unpopMatrix，且两者不相交。电影评分的 MAE 值如图 3 所示。可以看出，邻域用户数 k 值的变化对算法的推荐性能具有一定的影响，当用户数量过多或过少时都可能会降低推荐的准确性。若邻域用户 k 值太小，则协同过滤过程中的评分数减少，降低了该推荐算法得到的预测评分的可信性。若邻域用户 k 值太大，则可能会融入一些伪评价噪声，因此也降低了评价预测准确性。实验结果表明，基于欧氏距离的 APW 算法的邻域用户数量 $k=7$ 时效果最好。在基于曼哈顿距离的 APW 算法中，MAE 随着 k 值的增大而增大。

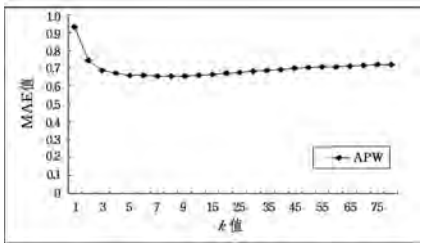

图 3 邻域用户数量对算法推荐性能的影响

3.2 性能对比

为了进一步验证所提算法的性能，将其与现有的基于用户相似度的协同过滤算法和基于用户的改进的协同过滤算法进行比较，即文献[9]算法、文献[10]算法以及 cosine 法和 person 法。在已有的 Movielens 数据集上进行测试，不同邻域用户数量下的 MAE 值如图 4、图 5 所示。可以看出，MAE 随着邻域数量的变化而变化且不同算法的变化趋势不同，在各自最佳邻域数量下所获得的推荐 MAE 值中，提出的 APW 算法的 MAE 值最小，这进一步证明了所提算法的有效性。

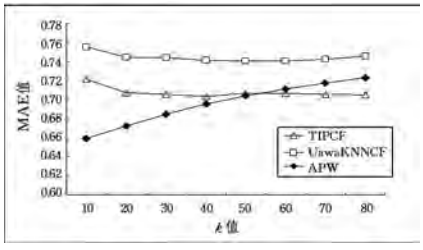

图 4 改进的协同过滤算法比较

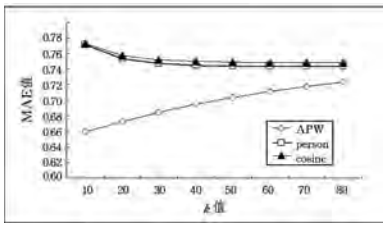

图 5 APW 算法与基于用户相似度的协同过滤算法的比较

结束语 针对传统个性化推荐方法的缺陷，为提高推荐准确性，提出一种结合数据集特征和新的加权法的协同过滤推荐算法。利用项目过滤分步实现项目-评分矩阵的构建。通过提出对用户加权的新算法来构建更为有效的协同过滤方法，并以此对项目给出最终的评分。在 Movielens 的电影推荐数据集上的实验结果表明，该方法能够准确地给出相应推荐。后续可将与算法结合的数据集划分程度和方式的不同对推荐权重进行进一步加权。

参 考 文 献

[1] 李容，李明奇，郭文强. 基于改进相似度的协同过滤算法研究[J]. 计算机科学，2016，43(12)：206-208.

[2] 徐蕾，杨成，姜春晓，等. 协同过滤推荐系统中的用户博弈[J]. 计算机学报，2016，39(6)：1176-1189.

[3] XUE G R，LIN C，YANG Q，et al. Scalable collaborative filtering using cluster-based smoothing[C]// International ACM Sigir Conference on Research & Development in Information Retrieval. 2005：114-121.

[4] 姚彬修，倪建成，于苹苹，等. 基于多源信息相似度的微博用户推荐算法[J]. 计算机应用，2017，37(5)：1382-1386.

[5] 于洪涛，周倩楠，张付志. 基于项目流行度和新颖度分类特征的托攻击检测算法[J]. 工程科学与技术，2017，49(1)：176-183.

[6] SHI Y，LARSON M，HANJALIC A. Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering[C]// Proceedings of the Third ACM Conference on Recommender Systems. ACM，2009：125-132.

[7] BANDA L，BHARADWAJ K K. Evaluation of Collaborative Filtering Based on Tagging with Diffusion Similarity Using Gradual Decay Approach[M]// Advanced Computing，Networking and Informatics- Volume 1. Springer International Publishing，2014：421-428.

[8] 黄文明，莫阳. 基于文本加权 KNN 算法的中文垃圾短信过滤[J]. 计算机工程，2017，43(3)：193-199.

[9] 郑洁，钱育蓉，杨兴耀，等. 基于信任和项目偏好的协调过滤算法[J]. 计算机应用，201636(10)：2784-2788.

[10] 王吉源，黎晨，王婵娟. 用户属性加权活跃近邻的协同过滤算法[J]. 计算机应用研究，2016，(12)：3625-3629.

[11] SHANG M S，JIN C H，ZHOU T，et al. Collaborative filtering based on multi-channel diffusion [J]. Physica A Statistical Mechanics & Its Applications，2009，388(23)：4867-4871.

[12] 王伟，徐平平，王华君，等. 基于概率回归模型和 K-最近邻的电子商务个性化推荐方案[J]. 湘潭大学自科学报，2016，38(1)：97-100.

[13] BOBADILLA J，ORTEGA F，HERNANDO A. Recommender systems survey [J]. Knowledge-Based Systems，2013，46(1)：109-132.

[14] 徐雅斌，孙晓晨. 位置社交网络的个性化位置推荐[J]. 北京邮电大学学报，2015，38(5)：118-124.