文章编号：2095-2163(2021)09-0161-05 中图分类号：TP391 文献标志码：A

结合物品属性权重的混合推荐算法

马梦馨，王国中
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：自推荐系统被提出以来，各类算法层出不穷，各有利弊。数据稀疏性和冷启动问题是大部分推荐算法存在的缺点，将各类推荐算法混合，扬长避短，能很好的解决这些问题，传统的混合算法是将几种方法进行简单的线性组合。本文将物品属性权重引入相似性计算，再将改进的余弦相似性与之结合，生成一种动态的计算物品相似度的算法，将基于物品的协同过滤和基于内容的推荐的算法进行结合。实验数据表明该算法提高了推荐准确性的同时，还有效缓解了数据稀疏性和冷启动问题。

关键词：协同过滤；物品属性；混合算法；相似性

Hybrid recommendation algorithm combined with item attribute weight

MA Mengxin, WANG Guozhong
(School of Electronic and Electrical Engineering ,Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 All kinds of algorithms have their own advantages and disadvantages. Data sparsity and cold start problems are the shortcomings of most recommendation algorithms. Though mixing all kinds of recommendation algorithms can solve these problems, the traditional hybrid algorithm is a simple linear combination of several methods. In this article, attribute weight is introduced into the similarity calculation and the improved cosine similarity is combined with it to generate a dynamic algorithm to calculate the similarity of the item, which combines the collaborative filtering based on the item and the recommendation based on the content. Experimental data shows that the algorithm improves the accuracy of recommendation and at the same time it also effectively alleviates the problems of data sparsity and cold start.

【Key words】 collaborative filtering; item attributes; hybrid algorithm; similarity

0 引 言

随着信息技术和互联网技术的发展，互联网提供的平台和数据越来越多，而不同的人兴趣爱好截然不同，越来越难以从大量的信息中找到自身感兴趣的信息，信息也越来越难展示给可能对其感兴趣的用户，推荐系统应运而生。推荐系统本质上是在用户需求不明确的情况下，从海量信息中为用户寻找有用信息的技术手段。经过二十多年的发展，推荐系统被广泛应用于电子商务平台、新闻媒体领域以及广告的个性化推荐等。

目前市面上比较常用的推荐算法有协同过滤推荐算法（Collaborative Filtering Recommendation, CF），其中包括基于用户的协同过滤（User Based CF）和基于物品的协同过滤（Item Based CF），基于内容的推荐算法（Content-Based Recommendation, CB）和混合推荐算法（Hybrid Recommendation, HR）等。

协同过滤推荐算法在一般情况下表现良好，但是在有新用户或新物品加入时，由于没有历史数据，所以无法进行推荐，存在冷启动和数据稀疏性问题。Liu 等人提出在传统矩阵分解模型的基础上，通过整合多关系社交网络的用户偏好，获得信任和信任功能矩阵，有效缓解了数据稀疏性问题[1]；Yan 等人提出了将 Jaccard 相似性计算方法用于基于多层感知机的电影推荐模型，解决数据稀疏性问题[2]；苑等人根据社交活动提出一种新的用户相似度计算方法来提高推荐精度[3]；过等人改进了奇异值分解（SVD）算法和二分 K-均值聚类算法，解决协同过滤算法稀疏性较大和扩展性较差的问题[4]。

基于内容的推荐算法不存在冷启动问题，但是存在提取特征困难、无法挖掘用户的潜在兴趣等缺点。王等人将项目粒度化，用户信息生成用户粒度序列来提取特征，提高推荐精度[5]。

基金项目：国家重点研发计划资助（2019YFB1802700）。

作者简介：马梦馨（1997-），女，硕士研究生，主要研究方向：知识图谱、推荐系统；王国中（1962-），男，博士，教授，硕士生导师，主要研究方向：数字音视频信息处理、智能信息处理、推荐系统。

收稿日期：2021-04-20

混合推荐算法能根据不同的方式将多种算法相结合,扬长避短,提高推荐精度,解决冷启动和数据稀疏等问题。刘等人将不同用户对于不同物品的个性化行为特征指数引入到相似度的计算中,动态计算权重,提高混合推荐算法的推荐效果[6];Fan 等人采用分类和聚类算法来挖掘项目和用户的历史数据,改进混合推荐算法,解决电子商务推荐系统的问题[7];李等人考虑了用户评分尺度及用户活跃度对物品相似性的影响,动态生成权重因子,提高推荐精度[8];随着深度学习的发展,田等人提出了一种基于隐狄利克雷分布(LDA)与卷积神经网络(CNN)的概率矩阵分解推荐模型(LCPMF),获取深层项目特征,提高推荐精度[9]。

本文在传统的混合推荐模型的基础上,引入物品属性的权重,改进了相似性计算方法,将协同过滤推荐算法与基于内容的推荐算法动态结合,解决冷启动和数据稀疏性问题,提高推荐精度。

1 相关算法理论

1.1 评分矩阵

定义推荐系统中 $U=\{u_1,u_2,\cdots,u_m\}$ 为所有 m 个用户的集合,$I=\{i_1,i_2,\cdots,i_n\}$ 为所有 n 个物品的集合,两个集合组成了一个 $M\times N$ 的矩阵,此矩阵为用户-物品评分矩阵。见表 1,矩阵中 r_{ui} 为用户 u 对物品 i 的评分,若 r_{ui} 为 0,则说明用户对该物品没有评分,评分值越高说明用户对该物品越感兴趣。

表 1 用户-物品评分矩阵

Tab. 1 User-item scoring matrix

	i_1	i_2	…	i_n
u_1	r_{11}	r_{12}	…	r_{1n}
u_2	r_{21}	r_{22}	…	r_{2n}
⋮				⋮
u_m	r_{m1}	r_{m2}	…	r_{mn}

1.2 相似性计算

推荐算法中,常用的计算方法有欧氏距离、余弦相似度和修正的余弦相似度等,使用场景各不相同。

欧氏距离是衡量同一空间下两个点,度量的是两个点的绝对差异,适用于分析用户的能力模型,定义如式(1):

$$E(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \tag{1}$$

余弦相似度度量的是两个向量之间的夹角,其在度量文本相似度、用户相似度、物品相似度时较为常用。定义如式(2):

$$\mathrm{COS}(p,q)=\frac{\sum_{i=1}^{n}(p_i*q_i)}{\sqrt{\sum_{i=1}^{n}p_i^2}*\sqrt{\sum_{i=1}^{n}q_i^2}} \tag{2}$$

修正的余弦相似度是将数据中心化后再求余弦相似度,定义如式(3):

$$\mathrm{ACOS}(p,q)=\frac{\sum_{i=1}^{n}(p_i-\bar{p})*(q_i-\bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i-\bar{p})^2}*\sqrt{\sum_{i=1}^{n}(q_i-\bar{q})^2}} \tag{3}$$

2 结合物品属性权重的混合推荐算法

2.1 物品流行度对相似性的影响

一般来说,热门物品会被用户喜欢的可能性大,但并不能说明用户的兴趣相同,热门物品对计算用户的相似性贡献不大,两个用户对冷门物品采取过同样的行为更能说明其兴趣度相同,二者更为相似,因此引入惩罚因子 θ_i 惩罚用户 u_1、u_2 共同兴趣列表中热门物品对其相似度的影响,θ_i 的公式定义如式(4):

$$\theta_i=\frac{1}{\log(1+|N(i)|)} \tag{4}$$

其中,$N(i)$ 表示对物品 i 有过评分的用户集合。

引入惩罚因子后的相似度为计算公式(5):

$$sim^{itemcf}(i,j)=\frac{\sum_{i\in I_{u,v}}\theta_i*(R_{u,i}-\bar{R}_u)(R_{v,i}-\bar{R}_v)}{\sqrt{\sum_{i\in I_{u,v}}(R_{u,i}-\bar{R}_u)}*\sqrt{\sum_{i\in I_{u,v}}(R_{v,i}-\bar{R}_v)}} \tag{5}$$

其中,$I_{u,v}$ 表示用户 u 和用户 v 共同评过分的物品集合;$R_{u,i}$、$R_{v,i}$ 分别表示用户 u 和用户 v 对物品 i 的评分;$\bar{R}_u$ 和 $\bar{R}_v$ 分别表示用户 u 和用户 v 对所有物品评分的平均分。

2.2 物品属性相似性

基于内容的推荐算法是通过抽取物品本身的特征信息,形成关键词向量,然后与用户喜好特征向量进行相似度计算,将物品推荐给用户,通常用于文本推荐。

把一个物品看作一个文档,定义所有的文档集合为 $D=\{d_1,d_2,\cdots,d_t\}$,文档中的关键词集合定义为 $T=\{t_1,t_2,\cdots,t_s\}$,最终需要用一个向量表示一

个文档，定义 $\boldsymbol{d}_i = (\omega_{1,i}, \omega_{2,i}, \cdots, \omega_{s,i})$ 为物品 i 的关键词向量，其中 ω_{ni} 表示第 n 个词在文档 i 中的权重，数值越大表示越重要。定义好之后通常用词频-逆文档频率（TF-IDF）来表示文档，其定义如式（6）：

$$TF - IDF(t_k, d_i) = TF(t_k, d_i) * \log \frac{N}{n_k} \quad (6)$$

其中，$TF(t_k, d_i)$ 表示第 k 个词在文档 i 中出现的次数，n_k 是所有文档中包含第 k 个词的文档数量，最终第 k 个词在文档 i 中的权重如式（7）所示：

$$\omega_{k,i} = \frac{TF - IDF(t_k, d_i)}{\sqrt{\sum_{j=1}^{T} TF - IDF(t_j, d_i)}} \quad (7)$$

得到文档的特征向量权重之后，使用余弦相似度，得到文档之间的相似度，相似度定义如式（8）：

$$sim^{itemcb}(i,j) = \frac{\sum_{T_{i,j}} \omega_{k,i} * \omega_{k,j}}{\sqrt{\sum_{T_{i,j}} \omega_{k,i}} * \sqrt{\sum_{T_{i,j}} \omega_{k,j}}} \quad (8)$$

其中，$T_{i,j}$ 表示两文档之间共有的关键词。

2.3　混合模型相似性度量方法

通常协同过滤推荐算法效果优于基于内容的推荐算法，但是当新的用户或者物品加入时，系统就无法很好的进行推荐，且当用户物品矩阵极度稀疏时，计算出来的物品相似度可信度也不高，而基于内容的推荐算法能在一定程度上缓解物品冷启动问题，并且基于内容的推荐算法只考虑物品的属性，与用户的评价行为无关，能缓解数据稀疏性问题，所以将协同过滤算法中的相似性计算与物品属性相结合能缓解冷启动和数据稀疏性问题。

本文引入 λ 将两种相似性进行线性组合，由上文分析可知，当用户-物品矩阵极度稀疏时，使用基于内容的推荐算法要优于协同过滤推荐算法，所以定义 λ 的公式如式（9）：

$$\lambda = \frac{|U_i \cap U_j|}{|U_i \cup U_j|} \quad (9)$$

其中，U_i、U_j 表示对物品 i 和物品 j 评分的用户数；$U_i \cap U_j$ 表示对物品 i 和物品 j 共同评分的用户数；$U_i \cup U_j$ 表示物品 i 和物品 j 一共被多少用户评分。引入 λ 之后，将相似度计算公式进行线性组合，如式（10）所示：

$$sim^{item}(i,j) = \lambda sim^{itemcf}(i,j) + (1-\lambda) sim^{itemcb}(i,j) \quad (10)$$

由公式（10）可知，当存在冷启动问题或者用户-物品矩阵稀疏时，根据物品属性特征进行相似度计算的比重大；当数据稠密时，基于物品的协同过滤要优于基于内容的推荐，所以相似度计算时所占比重较大。这种线性结合的方式改善了推荐系统中的冷启动和数据稀疏性问题。

将混合的相似性计算方法引入到预测公式，得到用户 u 对物品 i 的评分预测公式（11）：

$$P_{u,i}^{item} = \bar{R}_i + \frac{\sum_{j \in M_i} sim(i,j) * (R_{u,j} - \bar{R}_j)}{\sum_{j \in M_i} |sim(i,j)|} \quad (11)$$

其中，M_i 为物品 i 的最近邻。

2.4　用户相似性

以上方法有效缓解了物品冷启动和数据稀疏性问题，但当新用户加入时，因为没有其历史行为记录，依然存在用户冷启动问题，只能根据用户自身的特征，为用户进行推荐。

影响用户喜好的特征主要有性别、年龄、职业、所在区域等信息，本文据此组成用户的内容向量，则用户 u 的特征集合为 $C_u = \{sex, age, occ, zip\}$，因为欧氏距离度量的是空间中两个点的绝对差异，所以本文使用欧氏距离，即公式（1）来计算用户之间的相似性。

冷启动用户的预测公式（12）为：

$$P_{u,i}^{user} = \frac{\sum_{v \in N_u} sim(u,v) * R_{v,i}}{\sum_{v \in N_u} |sim(u,v)|} \quad (12)$$

其中，N_u 表示用户 u 的最近邻。

2.5　推荐过程

为了解决数据稀疏性和冷启动问题，本文结合物品属性，将基于物品的协同过滤和基于内容推荐的相似性度量方法进行动态结合，形成一种新的相似性度量方法，解决物品冷启动和数据稀疏性问题，并且通过计算用户属性来解决用户冷启动问题。具体推荐过程如下：

Step 1　判断目标用户是否是冷启动用户，是则跳到 Step2，不是则跳到 Step3；

Step 2　冷启动用户的相似性计算，之后预测评分；

Step 3　非冷启动用户的相似性计算，评分预测；

Step 4　完成 Top-N 推荐。

3　实验数据及结果分析

3.1　数据集

为了验证本文算法的有效性，使用 MovieLens

1M 数据集,该数据集包含6 040个用户对3 900部电影的1 000 209条评分记录,数据稀疏度达 95.75%。将数据集按照 8:2 划分为训练集和测试集,数据集中用户的属性包括了用户的 ID、性别、年龄、职业 ID 和邮编等字段,电影的属性有电影 ID、电影名、电影年份和电影风格等。

3.2 评价指标

推荐系统中常用的评价标准有平均绝对误差(*MAE*)、均方根误差(*RMSE*)、准确率(*Precision*)和 *F* 值等,本实验采用 *MAE* 作为度量标准,其定义为式(13):

$$MAE = \frac{\sum_{i=1}^{n} |p_{u,i} - r_{u,i}|}{n} \tag{13}$$

其中,$p_{i,j}$ 表示用户 u 对物品 i 的预测评分;$r_{u,i}$ 表示用户 u 对物品 i 的实际评分;n 为数据集中记录评分的个数。

MAE 计算的是真实值与预测值之间的差异,数值越小说明准确性越高。

3.3 实验结果

通过实验测得本文算法在不同 N 的取值下的绝对误差,见表 2。由表 2 可知,N 取值在[10,60]范围内,精确性逐渐升高。

表 2 算法在不同 *N* 的取值下的平均绝对误差

Tab. 2 The average absolute error of the algorithm at different *N* values

N	10	20	30	40	50	60
MAE	0.7	0.685	0.683	0.679	0.675	0.655

3.3.1 算法推荐精准度比较

为了验证本文算法的优化效果,本文选取改进的基于物品的协同过滤、基于内容的推荐算法与本算法进行对比实验,分别设置不同最近邻值测试 *MAE* 值的大小,实验结果如图 1 所示。可以看出本文提出的推荐算法无论 N 取何值,效果都远大于基于物品的协同过滤和基于内容的推荐。

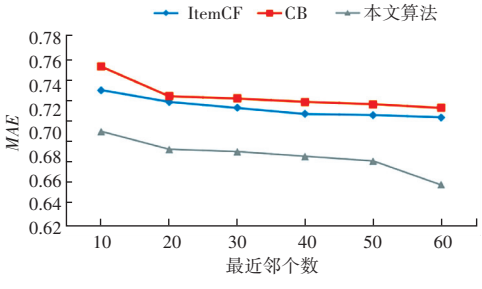

图 1 推荐准确度对比

Fig. 1 Recommended accuracy comparison

3.3.2 算法缓解数据稀疏性能力的比较

为了测试本文算法解决数据稀疏性问题的能力,本实验的最近邻数确定为 60,并且在数据集中随机删除部分数据,改变评分矩阵的稀疏性再次进行对比实验,测试算法效果,实验结果如图 2 所示。

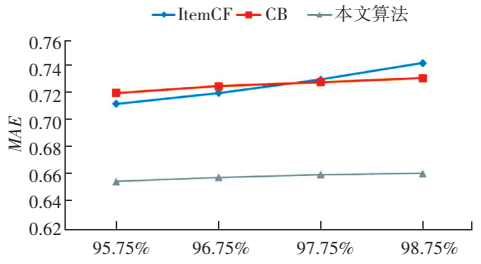

图 2 数据稀疏性对比

Fig. 2 Data sparsity comparison

由图 2 可知基于内容的推荐算法在数据极度稀疏情况下算法效果要优于协同过滤推荐算法,而本文提出的算法在数据稀疏的情况下,效果要明显优于其它两种算法,有效缓解了数据稀疏性的问题。

3.3.3 算法缓解冷启动能力的比较

本实验用来验证算法解决冷启动问题的能力,在测试集中抽取 100 个物品作为新物品,100 个用户作为新用户,将训练集中对应的 100 个物品和用户的评分记录置为 0,使用新的训练集和测试集进行实验。本实验将基于内容的推荐算法作为对比,结果如图 3 所示。

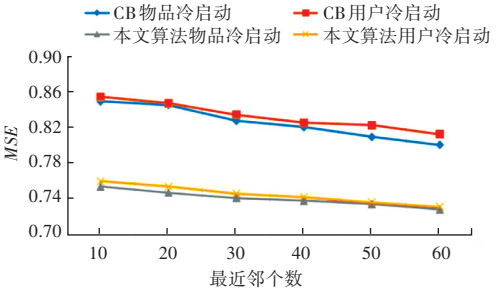

图 3 冷启动问题对比

Fig. 3 Cold start problem comparison

由图 3 可知,不管是用户冷启动还是物品冷启动,本文算法的精确性都远高于基于物品的协同过滤算法,实验表明,本算法能有效缓解冷启动问题。

4 结束语

本文对传统的混合推荐算法进行了优化,结合物品属性特征权重改进了相似度度量方法,并根据用户-物品矩阵稀疏性的差异,自适应的调整不同算法的相似性计算方法所占的比重,极大地提高了

(下转第 169 页)