基于重叠度和双重属性的协同过滤推荐算法

张 博 刘学军 李 斌
(南京工业大学电子与信息工程学院 南京 211816)

摘 要 协同过滤是现行推荐系统中应用最广泛也是最成功的推荐技术之一,然而传统的协同过滤推荐算法存在着邻居选取片面性和推荐精度低的问题。针对上述问题,提出了一种基于重叠度和双重属性的协同过滤推荐算法。首先基于相似度和重叠度的共同计算结果选取推荐对象集;然后提出了双重属性的概念,分别计算推荐用户的信任度和目标项目的受欢迎度;最后兼顾两个群体,根据用户和项目两方面的评分信息完成对目标用户的推荐。实验结果证明该算法较传统的协同过滤推荐算法在邻居选取和推荐质量方面均有显著的提高。

关键词 重叠度,推荐对象集,双重属性,协同过滤,推荐系统
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.048

Collaborative Filtering Recommendation Algorithm Based on Overlap Degrees and Double Attributes

ZHANG Bo LIU Xue-jun LI Bin
(College of Electronic and Information Engineering, Nanjing Tech University, Nanjing 211816, China)

Abstract Recently, collaborative filtering is one of the most widely used and successful recommendation technology in the recommender system. However, the traditional collaborative filtering recommendation algorithm has the disadvantages of the one-sidedness for selecting neighbors and lower recommendation precision. In order to solve the problems, this paper proposed a collaborative filtering recommendation algorithm based on overlap degrees and double attributes. Firstly, we used the aggregated results of similarity and overlap degrees to select the recommended set of objects. Then, we proposed the concept of double attributes, and calculated the reliability of the target user and the popularity of the target item respectively. At last, taking into account the two groups, we used both user and item rating information to generate recommendation for the target user. Experimental results show the proposed algorithm is improved significantly in terms of neighbor selection and recommendation quality compared to traditional collaborative filtering recommendation algorithm.

Keywords Overlap degrees, Recommender object, Double attributes, Collaborative filtering, Recommender systems

1 引言

随着 Web2.0 技术的迅速发展,用户主动获取与分享的积极性不断提高,网络上的信息量急剧增长,用户逐渐陷入选择信息的困窘中。推荐系统(Recommender Systems)通过收集和分析用户的各种数据为用户推荐其所需要的信息和服务,逐渐成为解决互联网上信息过载问题的有效工具。各种推荐技术越来越多地被应用到各种门户网站和电子商务系统之中,例如 Amazon、eBay、阿里巴巴等。

目前主要的推荐技术包括协同过滤(Collaborative Filtering)推荐、基于内容的推荐、基于知识的推荐以及混合推荐等,其中协同过滤技术是目前推荐系统中所采用的最为重要也是最成功的技术之一。其认为兴趣相似的用户会喜欢相似的项目,利用用户对项目的评分信息,寻找与目标用户兴趣相似的用户进而为目标用户推荐相关项目。协同过滤推荐主要包括基于内存和基于模型两大类;前者直接根据用户或项目之间的相似度进行推荐,后者通过训练内存中的数据得到数据模型而后进行推荐。

传统的协同过滤推荐算法在选取目标对象的邻居时通常仅仅依靠对象间的相似度,且在预测目标对象评分时只考虑到了对象群内个体与个体之间的相互作用,而忽略了另一个对象关联群的影响作用,从而导致推荐结果的片面性。针对上述问题,本文从重叠度和用户与项目双重属性的思路出发,提出了一种新的基于重叠度和双重属性的协同过滤推荐算法(Collaborative Filtering recommendation algorithm based on Overlap Degrees and Double Attributes, CF-ODDA)。

本文第 2 节介绍相关工作;第 3 节介绍传统的协同过滤推荐算法(CF);第 4 节介绍 CF-ODDA 的具体实现;第 5 节给出实验和结论;最后对全文进行总结和展望。

2 相关工作

为了提高系统推荐精度,改善传统的协同过滤推荐算法

到稿日期:2015-03-04 返修日期:2015-06-07 本文受国家自然科学基金(61203072),国家公益性科研专项(201310162)资助。
张 博(1991-),女,硕士生,主要研究方向为数据挖掘、社会推荐,E-mail:zhangbo_hello@163.com;刘学军(1971-),男,博士,教授,CCF 高级会员,主要研究方向为数据库、数据挖掘、传感器网络等;李 斌(1979-),男,硕士,讲师,主要研究方向为传感器网络、智能信息处理。

中存在的数据稀疏性、冷启动和可扩展性等不足，国内外学者进行了一系列研究，提出了诸多改进的启发式推荐算法。例如，贾冬艳等人[1]提出了一种基于双重邻居选取策略的协同过滤推荐算法，以目标用户对兴趣相似用户的信任度作为选取可信邻居用户的第二重依据，提高了系统推荐精度，而且具有较强的抗攻击能力，但对于冷启动用户的准确性还有待提高。Yamashita 等人[2]将基于用户的协同过滤和基于项目的协同过滤融合在一起，提出了一种新的推荐方法，较好地解决了数据稀疏性问题。Wen 等人[3]提出了一种改进的基于聚类方法的项目协同过滤算法，在推荐过程中利用了聚类技术，在一定程度上克服数据稀疏性影响的同时也提高了推荐结果的准确性。

除此之外，许多学者提出了利用潜在因子模型来解决相关问题。在潜在因子模型中，一个用户 u 对应一个用户因子向量 p_u，一个项目 i 对应一个项目因子向量 q_i，最后通过内积函数 $\hat{r}_{ui}=b_{ui}+p_u^{\mathrm{T}}q_i$ 计算预测值。例如，Zhang 等人[4]利用灵活的回归先验探索矩阵分解，任何新的回归过程都可以通过标准的回归路线插入到本模型拟合过程中的一些中间步骤来。Gao 等人[5]提出通过区分用户在不同时间段的潜在因素，并利用多种策略将用户在不同时间段的潜在因素聚集在一起，来为社会网络中的用户推荐位置。Zhou 等人[6]提出了一种函数分解法来处理冷启动问题。Ma 等人[7]使用集体概率因子模型来进行网站推荐。Wang 等人[8]在网站推荐中运用了概率模型，实现了个性化排序。然而，虽然潜在因子模型能够在一定程度上缓解数据稀疏性带来的问题，但在降维的同时也会导致相关信息的丢失，从而产生不准确的推荐结果。

近年来，在推荐过程中引入社会网络信息成为了学者们研究的热点之一。Yang 等人[9]提出了一种基于朋友圈的社会化推荐算法，认为用户受到其所处的朋友圈的影响，在不同的朋友圈中，用户间的信任关系也是不同的。Noel 等人[10]在社会协同过滤中提出了一种新的目标函数，综合考虑了用户特征、用户间的信息传播和用户间兴趣的问题，并利用矩阵分解进行求解。郭磊等人[11]从推荐对象间关联关系的角度出发，假设具有关联关系的推荐对象更容易受到同一用户的关注，提出了一种结合推荐对象间关联关系进行推荐的算法。张斌等人[12]提出了一种基于标签系统中对象间关系与资源内容融合分析的标签推荐方法，通过一种新的概率潜在主题模型 TSM/Forc，为用户提供更加准确的推荐结果。Jiang 等人[13]提出了一种基于社会上下文信息的推荐算法，根据推荐对象间的内容相似性对目标函数进行约束，为用户推荐 tweets。Yu 等人[14]研究了个性化实体推荐，提出了为每个聚类专门建立一个本地检索模型来共享相关信息。上述研究分别从朋友关系、用户特征、用户间兴趣、推荐对象间关系这几个方面证实了社会网络信息能够在一定程度上解决推荐系统中存在的问题，提高推荐精度。

基于以上研究，本文提出了一种基于重叠度和双重属性的协同过滤推荐算法(CF-ODDA)，通过相似度和重叠度两方面选取推荐对象集，得到目标对象的最佳邻居集，同时在预测目标对象的评分值时，加入了用户和项目的双重属性，综合考虑两个群体的影响，最终达到拉近预测评分与实际评分间的距离和提高推荐精度的效果。

3 传统的协同过滤推荐算法(CF)

在协同过滤推荐系统中，用户对所有项目的评价数据库由 m 个用户的集合 $U=\{u_1,u_2,\cdots,u_m\}$ 和 n 个项目的集合 $I=\{i_1,i_2,\cdots,i_n\}$ 构成，用户-项目评分数据集用 $m\times n$ 阶矩阵 R 表示，其中 $R_{i,j}(1\leqslant i\leqslant m,1\leqslant j\leqslant n)$ 表示用户 u_i 对项目 i_j 的评分，如果用户 u_i 对项目 i_j 没有进行评分，则记作 $R_{i,j}=\emptyset$。本文以基于用户的协同过滤推荐算法为例。

3.1 邻居用户集的选取

在推荐过程中，选取与目标用户相似度较大的用户作为邻居是保障推荐质量的关键。在协同过滤算法中，传统的相似度计算方法主要有余弦相似性、改进的余弦相似性和 Pearson 相关系数 3 种。余弦相似性实现简单、速度快，但未体现出用户评分的统计特征。修正的余弦相似性更多地体现了用户的相关性而不是相似性。Pearson 相关系数则不仅考虑了项目的平均评分，且对于某用户总是倾向于给出比另一个用户更高的评分值，而二者的分值之差又始终保持一致，在这种情况下，能依旧认为他们具有很好的相似性。因此，在协同过滤算法中一般采用 Pearson 相关系数来计算用户间的相似度，确定邻居用户，其计算公式如下：

$$Sim_{a,b}=\frac{\sum\limits_{i_t\in I_{a,b}}(R_{a,t}-\overline{R_a})(R_{b,t}-\overline{R_b})}{\sqrt{\sum\limits_{i_t\in I_{a,b}}(R_{a,t}-\overline{R_a})^2}\sqrt{\sum\limits_{i_t\in I_{a,b}}(R_{b,t}-\overline{R_b})^2}} \tag{1}$$

其中，$Sim_{a,b}$ 表示目标用户 u_a 与用户集 U 中其他用户 $u_b(u_b\in U)$ 的相似度。$R_{a,t}$ 和 $R_{b,t}$ 分别表示 u_a 和 u_b 对项目 i_t 的评分。$\overline{R_a}$ 和 $\overline{R_b}$ 分别表示 u_a 和 u_b 的平均评分。$I_{a,b}$ 表示 u_a 和 u_b 的共同评分项目集。

将所有的相似度 Sim 从大到小排列为 $Sim_{a,1}$，$Sim_{a,2}$，$\cdots$，$Sim_{a,m}$，取前 k 个，则与此对应的 k 个用户就是目标用户 u_a 的邻居用户集，定义为 $S(u_a)$ 且 $|S(u_a)|=k$。如果 $Sim_{a,b}$ 排在前 k 名，则用户 u_b 就是目标用户 u_a 的邻居用户，即 $u_b\in S(u_a)$。

3.2 推荐结果的生成

在传统的协同过滤推荐算法中，首先根据用户相似性选取邻居集 $S(u_a)$，然后 $\forall u_b\in S(u_a)$ 作为推荐用户，针对项目集 $I_{a,b}=\{i_j\mid R_{a,j}=\emptyset,R_{b,j}\neq\emptyset,i_j\in I\}$ 中的每个项目，根据式(2)对目标用户 u_a 在目标项目 i_j 进行评分预测。

$$P_{a,j}=\overline{R_a}+\frac{\sum\limits_{u_b\in S(u_a)}Sim_{a,b}\times(R_{b,j}-\overline{R_b})}{\sum\limits_{u_b\in S(u_a)}Sim_{a,b}} \tag{2}$$

其中，$P_{a,j}$ 表示推荐用户 u_b 对目标用户 u_a 在目标项目 i_j 上的预测评分。$R_{b,j}$ 表示 u_b 对 i_j 的评分。$\overline{R_a}$ 和 $\overline{R_b}$ 分别表示 u_a 和 u_b 的平均评分。$Sim_{a,b}$ 表示 u_a 和 u_b 的相似度。

4 基于重叠度和双重属性的协同过滤推荐算法(CF-ODDA)

4.1 选取目标对象的推荐对象集

在传统的协同过滤算法中，我们通常依赖于 kNN 方法寻找 k 个邻居，但当某一个数据相当稀疏、信息量有限时，选择的 k 近邻中也会包含一些与目标相似度非常小的对象，那么最大 k 近邻也会产生一些不相似的个体进行协同过滤，降低 k 近邻的可靠性，从而导致推荐精度的下降。研究表明，对象

间的关系对用户的选择也有着很大的关系。因此，本文在考虑相似度的同时，借助两者之间关系的相似度，来综合考虑两方面结果选取目标对象的推荐对象集。

(1)对于用户来说，利用 Pearson 相关系数计算出目标用户与用户集中其他用户的相似度，而后兼顾社交圈重叠度[15]来选择推荐集。用户处在不同的社会关系中，亲人、朋友、同事等都是他们在不同社交圈下的不同角色，同一社交圈的人相互联系密切。在此环境下，如果两个用户的社交圈重叠度越高，则他们的相似度也就越高。记目标用户 u_a 的社交圈为 $u_ac_1, u_ac_2, \cdots, u_ac_p$，用户 u_b 的社交圈记为 $u_bv_1, u_bv_2, \cdots, u_bv_q$，则社交圈重叠度的计算公式如下：

$$CircleSim_{a,b}=\sum_{i=1}^{p}\sum_{j=1}^{q}OLD_{i,j} \tag{3}$$

$$OLD_{i,j}=S_{i,j}\times(T_{i,j}+1) \tag{4}$$

其中，$CircleSim_{a,b}$ 表示目标用户 u_a 与邻居用户 u_b 的社交圈重叠度。$OLD_{i,j}$ 表示社交圈 u_ac_i 与 u_bv_j 的重叠度。记由 u_ac_i 与 u_bv_j 重叠顶点组成的图为 $Circle_{i,j}$，则 $S_{i,j}$ 表示图 $Circle_{i,j}$ 中点的个数，$T_{i,j}$ 表示图 $Circle_{i,j}$ 中边的个数。

在此，我们规定对于每一位目标用户 u_a，有 $\sum_{u_b\in U}CircleSim^*_{a,b}=1$，则将目标用户 u_a 与邻居用户 u_b 的社交圈重叠度 $CircleSim_{a,b}$ 标准化为：

$$CircleSim^*_{a,b}=\frac{CircleSim_{a,b}}{\sum_{u_b\in U}CircleSim_{a,b}} \tag{5}$$

$\forall u_b\in U$，其与目标用户 u_a 的联合相似度 $CSim_{a,b}$ 定义为 Pearson 相关系数与社交圈重叠度的加权和，计算公式如下：

$$CSim_{a,b}=\alpha Sim_{a,b}+(1-\alpha)CircleSim^*_{a,b} \tag{6}$$

其中，α 表示权重($0\leqslant\alpha\leqslant1$)，取 $CSim_{a,b}$ 最大的 k 个对应用户作为目标用户的推荐集，定义为 $C(u_a)$ 且 $|C(u_a)|=k$。

(2)对于项目来说，利用 Pearson 相关系数计算出目标项目与项目集中其他项目的相似度，而后兼顾类别圈重叠度来选择推荐集。通常情况下，一个项目具有多个标签，一个标签即代表一个类别，则一个项目就隶属于多个类别，例如电影《十面埋伏》是武侠片的代表，同时也是一部爱情片。由此我们认为两个项目所属的类别圈重叠度越高，其相似度也就越高。记目标项目 i_t 的社交圈为 $i_tc_1, i_tc_2, \cdots, i_tc_p$，项目 i_s 的社交圈记为 $i_sv_1, i_sv_2, \cdots, i_sv_q$，则类别圈重叠度的计算公式如下：

$$CircleSim_{t,s}=\sum_{i=1}^{p}\sum_{j=1}^{q}OLD_{i,j} \tag{7}$$

$$OLD_{i,j}=S_{i,j}\times(T_{i,j}+1) \tag{8}$$

其中，$CircleSim_{t,s}$ 表示目标项目 i_t 与邻居项目 i_s 的类别圈重叠度。$OLD_{i,j}$ 表示类别圈 i_tc_i 与 i_sv_j 的重叠度。记由 i_tc_i 与 i_sv_j 的重叠顶点构成的图为 $Circle_{i,j}$，则 $S_{i,j}$ 表示图 $Circle_{i,j}$ 中的点的个数，$T_{i,j}$ 表示图 $Circle_{i,j}$ 中边的条数。

在此，规定对于每一个目标项目 i_t 有 $\sum_{i_s\in I}CircleSim^*_{t,s}=1$，则将目标项目 i_t 与邻居项目 i_s 的类别圈重叠度 $CircleSim_{t,s}$ 标准化为：

$$CircleSim^*_{t,s}=\frac{CircleSim_{t,s}}{\sum_{i_s\in I}CircleSim_{t,s}} \tag{9}$$

$\forall i_s\in I$，其与目标用户 i_t 的联合相似度 $CSim_{t,s}$ 定义为 Pearson 相关系数与类别圈重叠度的加权和，计算公式如下：

$$CSim_{t,s}=\alpha Sim_{t,s}+(1-\alpha)CircleSim^*_{t,s} \tag{10}$$

其中，α 表示权重($0\leqslant\alpha\leqslant1$)，取 $CSim_{t,s}$ 最大的 k 个对应的项目作为目标项目的推荐集，定义为 $C(i_t)$ 且 $|C(i_t)|=k$。

对于目标对象而言，其推荐对象集选取算法如下(以选取目标用户的推荐集为例)：

算法 1 选取目标对象的推荐对象集 SRN(u_a)

输入：目标用户 u_a，目标用户 u_a 的社交圈 $u_ac_1, u_ac_2, \cdots, u_ac_p$，参数 k，参数 α，用户集 U，用户-项目评分矩阵 R

输出：目标用户 u_a 的推荐对象集 $C(u_a)$

Begin

1)$C(u_a)\leftarrow\emptyset$，Sum$\leftarrow$0；

2)For each $u_i\in U$ do

3)$Sim_{a,i}\leftarrow$similarity(u_a, u_i)；

4)检测 u_i 的社交圈，计算 $OLD_{a,i}$；

5)$CircleSim_{a,i}\leftarrow\sum OLD_{a,i}$；

6)End for

7)Sum$\leftarrow\sum_{u_i\in U}CircleSim_{a,i}$；

8)For each $CircleSim_{a,i}$ do

9)$CircleSim^*_{a,i}\leftarrow\frac{CircleSim_{a,i}}{\sum_{u_i\in U}CircleSim_{a,i}}$；

10)End for

11)For each $u_i\in U$ do

12)$CSim_{a,i}\leftarrow\partial Sim_{a,i}+(1-\partial)CircleSim^*_{a,i}$；

13)End for

14)Sort $LSim_{a,i}$；

15)$C(u_a)\leftarrow$最大的 k 个 $LSim_{a,i}$ 所对应的 u_i；

16)Return $C(u_a)$

End

4.2 双重属性计算

定义 1(评价项目组 O) 对于目标用户 $u_a\in U$，$\forall u_i\in C(u_a)$，u_i 评价过的项目即是 O 中的一个元素，则目标用户 u_a 的评价项目组 $O(u_a)$ 表示为：

$$O(u_a)=\{i_j\,|R_{i,j}\in R, u_i\in C(u_a), i_j\in I\} \tag{11}$$

定义 2(推荐用户组 W) 对于目标项目 $i_j\in I$，$\forall u_i\in U$，若 u_i 向任意 $u_k(u_k\in U, i\neq k)$ 推荐过项目 i_j 记作 $W_{i_j\to u_k}$，u_k 即是 W 中的一个元素，则目标项目 i_j 的推荐用户组 $W(i_j)$ 表示为：

$$W(i_j)=\{u_k\,|u_k\in W_{i_j\to u_k}, i_j\in I, u_k\in U\} \tag{12}$$

定义 3(接受用户组 A) 对于目标项目 $i_j\in I$，$\forall u_k\in W(i_j)$，若 u_k 接受了项目 i_j 记作 $A_{u_k\to i_j}$，u_k 即是 A 中的一个元素，则目标项目 i_j 的接受用户组 $A(i_j)$ 表示为：

$$A(i_j)=\{u_k\,|u_k\in A_{u_k\to i_j}, i_j\in I, u_k\in W(i_j)\} \tag{13}$$

在式(2)预测评分计算过程中，虽然综合考虑了目标对象与推荐对象的相似性和最相似的 h 个对象与目标对象 u_a 平均评分 $\overline{R_a}$ 的偏差，但由于相似性计算和评分标准存在较大的不确定因素，系统推荐精度明显下降。在实际应用中，用户总是容易同时受到推荐用户的信任度与推荐项目的受欢迎度两方面的影响。我们假设用户 u_a 过去评价的总次数为 10，u_b 评价的总次数为 1，则用户更相信 u_a 的推荐。另一方面，如果同时向 10 个用户推荐项目 i_j 和 i_k，其中有 9 个用户接受了项目 i_j，而项目 i_k 只有 1 个用户接受，则用户更愿选择项目 i_j。综合考虑，用户更倾向于接受用户 u_a 推荐的项目 i_j。

因此，本节引入双重属性，兼顾两个群体的影响，在宏观上把握推荐用户的信任度与推荐项目的受欢迎度，推荐用户

的信任度越高，推荐项目越受欢迎，用户越愿意采纳，推荐结果越精确，大大提高了推荐系统的准确性。

(1)用户的信任度(Reliability)

对于推荐用户 u_i 而言，$\forall u_i \in C, u_i$ 的信任度与其评价过的项目数量紧密联系，我们认为评价过较多项目的用户比没有评价过或评价过较少项目的用户更值得信任。例如，在美国的经济政策问题上，我们更愿意关注奥巴马的动态而不是希拉里的动态。

设 $RE_{a,i}$ 表示目标用户 u_a 对推荐用户 u_i 的信任度，计算公式如下：

$$RE_{a,i}=\frac{|I_{u_i}|}{|O(u_a)|} \tag{14}$$

其中，$|I_{u_i}|$ 表示推荐用户 u_i 评价过的项目数量，$|O(u_a)|$ 表示集合 $O(u_a)$ 中项目的个数。用户的信任度越高，对目标用户的影响越大，目标用户越可能接受其推荐的项目。

(2)项目的受欢迎度(Popularity)

除了推荐用户的信任度，项目的受欢迎度也是影响用户选择的关键因素之一。在选择过程中，用户一般更愿意接受热门项目。为此，我们用项目的接受比例来表示项目的受欢迎程度，向用户推荐项目时，项目接受比例越高，受欢迎度也就越高，用户则更愿意采纳。

设 P_{i_j} 表示推荐项目 i_j 的受欢迎度，计算公式如下：

$$P_{i_j}=\frac{|A(i_j)|}{|W(i_j)|} \tag{15}$$

其中，$|W(i_j)|$ 表示项目 i_j 被推荐给用户的总次数，$|A(i_j)|$ 表示接受项目 i_j 的用户数目。

4.3 推荐算法 CF-ODDA

传统的协同过滤推荐算法通常根据相似度的大小选取前 k 个作为目标对象的邻居对象集，不能准确地反映目标对象与邻居对象间的匹配度。因此，本文在用 Pearson 相关系数计算相似度的基础上，加入重叠度选取出推荐对象集，得到目标对象的最佳邻居集，同时在预测目标对象的评分值时，加入用户和项目的双重属性，综合考虑了两个群体的影响，最终完成对目标对象的推荐。

基于重叠度和双重属性的协同过滤推荐算法 CF-ODDA 的核心思想如下。

1)针对目标项目 i_i，利用算法 SRN 分别动态选取目标用户 u_a 和目标项目 i_i 的推荐对象集 $C(u_a)$ 和 $C(i_i)$；

2)根据式(14)和式(15)，分别计算出 $C(u_a)$ 中用户的信任度和目标项目 i_i 的受欢迎度；

3)根据 $C(u_a)$ 和 $C(i_i)$ 中的评分信息和双重属性，利用式(16)计算出目标用户 u_a 在目标项目 i_i 上的预测评分。

$$\begin{aligned}P_{a,i}=&\lambda\times\Big[\overline{R_a}+\frac{\sum\limits_{u_k\in C(u_a)}CSim_{a,k}\times(RE_{a,k}+1)\times(R_{k,i}-\overline{R_k})}{\sum\limits_{u_k\in C(u_a)}CSim_{a,k}}\Big]+\\&(1-\lambda)\times\Big[\overline{R_i}+\frac{\sum\limits_{i_j\in C(i_i)}CSim_{i,j}\times(P_{i_j}+1)\times(R_{a,j}-\overline{R_j})}{\sum\limits_{i_j\in C(i_i)}CSim_{i,j}}\Big]\end{aligned} \tag{16}$$

其中，λ 表示调和参数($0\leqslant\lambda\leqslant1$)，$P_{a,i}$ 表示目标用户 u_a 对目标项目 i_i 的预测评分，$\overline{R_a}$ 和 $\overline{R_k}$、$\overline{R_i}$ 和 $\overline{R_j}$ 分别表示用户 u_a 和 u_k、项目 i_i 和 i_j 的平均评分，$C(u_a)$ 和 $C(i_i)$ 分别表示目标用户 u_a 和目标项目 i_i 的推荐对象集，$RE_{a,k}$ 表示目标用户 u_a 对推荐用户 $u_k(u_k\in C(u_a))$ 的信任度，P_{i_i} 表示目标项目 i_i 的受欢迎度，$CSim_{a,k}$ 和 $CSim_{i,j}$ 分别表示用户 u_a 和用户 u_k、项目 i_i 和项目 i_j 的联合相似度。

根据上述算法思想，给出基于重叠度和双重属性的协同过滤推荐算法(CF-ODDA)如下。

算法 2 协同过滤推荐算法 CF-ODDA

输入：用户-项目评分矩阵 R

输出：用户 u_a 对项目 i_i 的预测评分 $P_{a,i}$

Begin

1)$C(u_a)\leftarrow SRN(u_a), C(i_i)\leftarrow SRN(i_i)$;

2)For each $u_k\in C(u_a)$ and i_i do

3)$RE_{a,k}\leftarrow reliability(u_a,u_k)$

4)$P_{i_i}\leftarrow popularity(i_i)$;

5)End for

6)$P_{a,i}\leftarrow Predict\_CF\text{-}ODDA(u_a,i_i)$;

7)Return $P_{a,i}$;

End

5 实验结果及分析

5.1 数据集

本文实验数据采用了两个数据集：

(1)美国明尼苏达州立大学 Group Lens 研究小组提供的 Movie Lens 数据集(http://www.grouplens.org)。该数据集中记录了用户对自己看过的电影的评分信息，每个用户至少评价了 20 部电影，且评分范围为 1～5，"1"表示"poor"(不喜欢)，"5"表示"perfect"(非常喜欢)。本文从中选取了 943 个用户对 1682 部电影的大约 100000 次评分作为实验数据集，大小为 100kB，稀疏度为 93.7%。

(2)雅虎实验室提供的 Yahoo! Music 数据集(http://webscope.sandbox.yahoo.com/)。该数据集中记录的是用户对音乐的评分信息。本文选取了用户对自己选择和系统推荐的音乐的评分作为实验数据集，包括 15400 个用户对 1000 首歌曲的大约 354000 次评分，大小为 1.2MB，稀疏度为97.7%。

实验中从每个数据集中随机抽取 80%作为训练集，其余的 20%作为测试集。

5.2 评价指标

本文采用平均绝对偏差 MAE(Mean Absolute Error)作为度量标准，它是一种统计精度度量方法，同时也是最常用的一种推荐质量度量方法。其通过计算预测评分与实际评分之间的偏差来衡量预测的准确性。MAE 的值越小，表明推荐质量越高。

假设预测的用户评分集合为 $\{p_1,p_2,\cdots,p_N\}$，对应的实际用户评分集合为 $\{q_1,q_2,\cdots,q_N\}$，N 表示预测的次数，则 MAE 的计算公式如下：

$$MAE=\frac{\sum\limits_{i=1}^{N}|p_i-q_i|}{N} \tag{17}$$

5.3 权重 α 的取值实验

本文在计算目标用户与其他用户的联合相似度时引入了权重 α，通过 α 调节 Pearson 相关系数与重叠度在联合相似度中所占的比值，以获得最优相似度结果，选取目标对象的最佳推荐对象集。因此选择恰当的权重 α 对协同过滤起着至关重要的作用。以权重 α 的取值为横坐标。实验结果如图 1 所示。

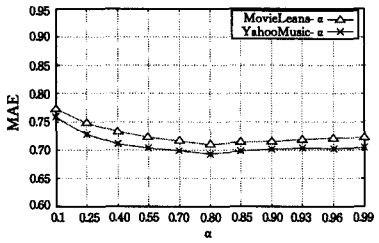

图 1　两个不同数据集上不同权重 α 取值结果的比较

从图中可以看出，当 $\alpha=0.8$ 时，MAE 值最小，推荐结果最优。当 $\alpha=0.1$ 即较小时，其 MAE 值较大。随着 α 逐步增加至 0.8 时，MAE 值逐步减小。权重 α 继续逐步递增，并越来越趋向于 1 时，MAE 值又逐步缓慢递增，并始终高于 $\alpha=0.8$ 并低于 $\alpha=0.1$ 时的 MAE 值。通过实验，可以认为权重 α 在协同过滤中起着重要作用，只有选择恰当的 α 才能获取最佳推荐对象集，得到最优推荐结果，MAE 才能降到最低。在接下来的实验中，将权重 α 的值均设置为 0.8。

5.4　选取预测目标的推荐对象方法的比较

本实验将文中 4.1 节提出的选取目标对象的推荐对象集 SRN 方法(Select Recommending Neighbor)与 DSN 方法[16]、传统的 kNN 方法进行了实验比较。通过选择较佳的推荐对象，为更好地进行协同过滤推荐打下基础。DSN 方法在考虑用户相似度的同时考虑了两者之间共同评价产品的个数，来选取目标对象的推荐对象。以选择的推荐对象个数为横坐标，以 MAE 为纵坐标，推荐对象个数从 10 开始，逐次增加，直至 100，并分别在两个不同数据集上验证，实验结果如图 2、图 3 所示。

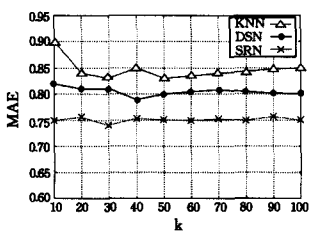

图 2　在 Movie Lens 数据集上选取推荐对象的不同方法比较

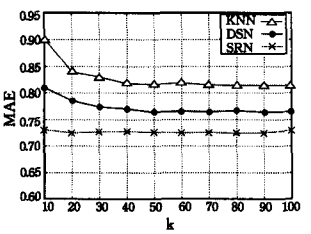

图 3　在 Yahoo! Music 数据集上选取推荐对象的不同方法比较

从图中可以看出，随着推荐对象个数的增加，MAE 值逐次减小，且更趋于稳定化。与此同时，无论采用哪种数据集，SRN 方法的 MAE 值都明显小于 DSN 方法和 kNN 方法。另外，对比两个数据集上的实验结果发现，随着训练用户数目的增多，目标用户的合适推荐对象就越容易找到，MAE 值也越小，这意味着通过更多用户的训练结果，得到的推荐结果也更好。由此我们认为选取目标对象的推荐对象集 SRN 方法能够有效地给用户推荐。在接下来的实验中，基于用户和基于项目的协同过滤推荐算法均采用选取目标对象的推荐对象集 SRN 方法进行实验。

5.5　调和参数 λ 的比率实验

本文在预测评分阶段引入调和参数 $\lambda(0\leqslant\lambda\leqslant1)$，目的是在进行协同过滤推荐时，通过 λ 平衡基于用户与基于项目之间的推荐结果。因此，本实验的目的是考察应该如何选择适当的调和参数 λ，以得出最佳的推荐结果。将基于用户的协同过滤和基于项目的协同过滤作为参考算法进行比较。以调和参数 λ 的取值为横坐标。实验结果如图 4、图 5 所示。

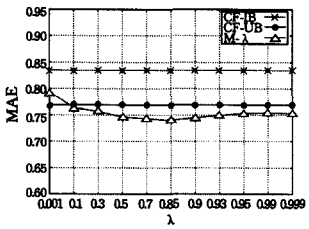

图 4　Movie Lens 数据集上不同调和参数 λ 的结果比较

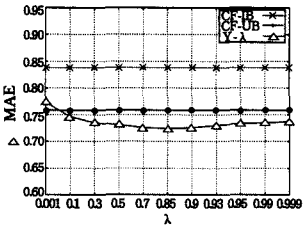

图 5　Yahoo! Music 数据集上不同调和参数 λ 的结果比较

从图中可以看出，当 $\lambda=0.85$ 时，MAE 值最小，CD-OPPA 推荐结果最优。观察其他取值，当 $\lambda=0.01$ 即较小时，其 MAE 值比 CF-IB 和 CF-UB 都大，但随着 λ 的逐步增加至 0.85时，MAE 值逐步减小，并取得比 CF-UB 和 CF-IB 更低的值，说明此时 CF-ODDA 推荐结果优于 CF-UB 和 CF-IB 推荐方法。同时，调和参数 λ 继续逐步递增，并越来越趋向于 1 时，其推荐结果仍优于其它两种参考算法。通过实验，我们认为调和参数 λ 在推荐过程中起着重要作用，只有选择适当的 λ 才能得到最好的推荐结果，MAE 才能降到最低。

5.6　CF-ODDA 算法与其他协同过滤算法的比较

本实验的目的是评价 CF-ODDA 算法的推荐精度，将本文提出的基于重叠度和双重属性的协同过滤推荐算法(CF-ODDA)与传统的协同过滤推荐算法(CF)进行了实验对比。另外，我们还对仅采用用户信任度(Collaborative Filtering recommendation algorithm based on User's Reliability, CF-UR)的推荐效果和仅采用项目受欢迎度(Collaborative Filtering recommendation algorithm based on Items' Popularity, CF-IP)的推荐效果进行了对比。实验结果如图 6、图 7 所示。

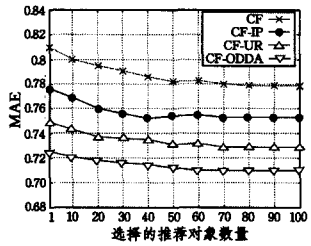

图 6　Movie Lens 数据集上 CF-ODDA 算法与其他协同过滤算法的比较

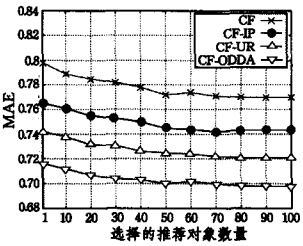

图 7　Yahoo! Music 数据集上 CF-ODDA 算法与其他协同过滤算法的比较

从图中可以看出，CF-UR 和 CF-IP 的推荐 MAE 值都明显小于 CF。由此可见，单独采用用户信任度或项目受欢迎度取得的推荐效果都优于 CF，说明了将用户信任度或项目受欢迎度引入到推荐过程中，可以改善系统推荐质量。另外，通过 CF-ODDA 的推荐 MAE 值与 CF-UR 和 CF-IP 的推荐 MAE 值对比发现，CF-ODDA 的推荐 MAE 值最小。因此，可以得到结论，同时使用用户信任度和项目受欢迎度的推荐算法比单独使用其中一种的推荐算法更有效果，即本文提出的基于重叠度和双重属性的协同过滤推荐算法 CF-ODDA 取得的效果最优。

结束语　随着互联网的迅速发展，传统的协同过滤算法在其推荐精度上已不能满足电子商务应用的需求。为此本文提出了一种基于重叠度和双重属性的协同过滤推荐算法，从

选取目标用户的推荐对象集出发，结合推荐用户的信任度和目标项目的受欢迎度完成对目标用户的推荐，并给出了 SRN 算法和 CF-ODDA 算法的具体实现过程。上述算法有效地改善了邻居选取片面性的问题，同时也提高了算法的推荐精度。

在协同过滤推荐算法中，冷启动也是其中一个主要问题。当某个新项目没有被任何用户评价过或新用户最初向系统提供的自己的信息非常有限时，推荐系统就无法发挥其作用。此时，我们可以考虑运用跨平台机制，整合该用户或项目在各个平台上的信息，根据其进行推荐。因此，怎样缓解协同过滤中的冷启动问题将是我们下一步的研究工作。

参 考 文 献

[1] Jia Dong-yan, Zhang Fu-zhi. A collaborative filtering recommendation algorithm based on double neighbour choosing strategy [J]. Journal of Computer Research and Development, 2013, 50 (5): 1076-1084(in Chinese)
贾冬艳，张付志. 基于双重邻居选取策略的协同过滤推荐算法[J]. 计算机研究与发展，2013，50(5)：1076-1084

[2] Yamashita A, Kawanura H, Suzuki K. Adaptive fusion method for user-based and item-based collaborative filtering [J]. Advances in Complex Systems, 2011, 14(2): 133-149

[3] Wen Jun-hao, Zhou Wei. An improved item-based collaborative filtering algorithm based on clustering method [J]. Journal of Computational Information Systems, 2012, 8(2): 571-578

[4] Zhang Liang, Deepak A, Chen Bee-chung. Generalizing matrix factorization through flexible regression priors[C]// Proceedings of the 5th ACM Conference on Recommender Systems. Chicago, USA, 2011: 13-20

[5] Gao Hui-ji, Tang Ji-liang, Hu Xia, et al. Exploring temporal effects for location recommendation on location-based social networks[C]// Proceedings of the 7th ACM Conference on Recommender Systems. New York, USA, 2013: 93-100

[6] Zhou Ke, Yang Shuang-hong, Zha Hong-yuan. Functional matrix factorizations for cold-start recommendation [C]// Proceedings of the 34th Int ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 315-324

[7] Ma Hao, Liu Chao, Irwin K, et al. Probabilistic factor models for Web site recommendation [C] // Proceedings of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. Beijing, China, 2011: 265-274

[8] Wang Hong-ning, He Xiao-dong, Chang Ming-wei, et al. Personalized ranking model adaptation for web search[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2013: 323-332

[9] Yang Xi-wang, Harald S, Liu Yong. Circle-based recommendation in online social networks [C] // Proceedings of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York, USA, 2012: 1267-1275

[10] Joseph N, Scott S, Khoi-Nguyen T, et al. New objective functions for social collaborative filtering[C]// Proceedings of the 21st International Conference on World Wide Web. Lyon, France: WWW, 2012: 859-868

[11] Guo Lei, Ma Jun, Chen Zhu-min, et al. Incorporating Item Relations for Social Recommendation[J]. Chinese Journal of Computers, 2014, 37(1): 219-288(in Chinese)
郭磊，马军，陈竹敏，等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报，2014，37(1)：219-288

[12] Zhang Bin, Zhang Yin, Gao Ke-ning, et al. Combining Relation and Content Analysis for Social Tagging Recommendation[J]. Journal of Software, 2012, 23(3): 476-488(in Chinese)
张斌，张引，高可宁，等. 融合关系与内容分析的社会标签推荐[J]. 软件学报，2012，23(3)：476-488

[13] Jiang Meng, Cui Peng, Liu Rui, et al. Social contextual recommendation[C]// Proceedings of the 21st ACM Conference on Information and Knowledge Management. Maui, USA, 2012: 45-54

[14] Yu Xiao, Ren Xiang, Sun Yi-zhou, et al. Personalized entity recommendation: A heterogeneous information network approach [C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014: 283-292

[15] Wang Yu, Gao Lin. Social circle-based algorithm for friend recommendation in online social networks[J]. Chinese Journal of Computers, 2014, 37(4): 801-808(in Chinese)
王珏，高琳. 基于社交圈的在线社交网络朋友推荐算法[J]. 计算机学报，2014，37(4)：801-808

[16] Huang Chuang-guang, Yin Jian, Wang Jing, et al. Uncertain neighbours' collaborative filtering recommendation algorithm[J]. Chinese Journal of Computers, 2010, 33(8): 1369-1377(in Chinese)
黄创光，印鉴，汪静，等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报，2010，33(8)：1369-1377

(上接第 218 页)

[5] Van der Aalst W M P, de Medeiros A K A, Weijters A. Genetic process mining [M] // Applications and Theory of Petri Nets 2005. Springer Berlin Heidelberg, 2005: 48-69

[6] Greco G, Guzzo A, Pontieri L. Mining hierarchies of models: From abstract views to concrete specifications [M] // Business Process Management. Springer Berlin Heidelberg, 2005: 32-47

[7] Fan Yu-shun. The basis of workflow management technology [M]. Beijing: Tsinghua University Press, 2001(in Chinese)
范玉顺. 工作流管理技术基础[M]. 北京：清华大学出版社，2001

[8] Wen Li-jie, Wang Jian-min, Sun Jia-guang. Modeling workflow patterns using coloured petri nets[J]. Computer Science, 2006, 33(6): 135-139(in Chinese)
闻立杰，王建民，孙家广. 用着色 Petri 网建模工作流模式[J]. 计算机科学，2006，33(6)：135-139

[9] Van der Aalst W M P. 过程挖掘：业务过程的发现、合规和改进[M]. 王建民，闻立杰，等译. 北京：清华大学出版社，2014

[10] Van der Aalst W M P. Service Mining: Using Process Mining to Discover, Check, and Improve Service Behavior[J]. IEEE Transactions on Services Computing, 2013, 6: 525-535

[11] Wang J, Wong R K, Ding J, et al. Efficient Selection of Process Mining Algorithms[J]. IEEE Transactions on Services Computing, 2013, 6: 484-496

[12] Van der Aalst W M P, Rubin V, Verbeek H M W, et al. Process mining: a two-step approach to balance between underfitting and overfitting [J]. Software & Systems Modeling, 2010, 9(1): 87-111