基于遗忘函数和领域最近邻的混合推荐研究①

朱国玮，周 利
（湖南大学工商管理学院，长沙 410082）

摘要：基于内容过滤和协同过滤是两大最为经典的推荐算法，但基于内容过滤存在新用户问题，没有考虑用户兴趣变化对推荐质量的影响，协同过滤则面临严峻的数据稀疏性和冷启动的挑战. 针对这些，提出混合推荐算法：基于非线性逐步遗忘函数建立用户兴趣模型，预测用户未评价商品评分；引入“领域最近邻”处理方法查找目标用户的最近邻，预测未评价商品评分，以此为基础做出推荐. 实验结果表明，本文方法能有效提高推荐质量.

关键词：混合推荐；非线性逐步遗忘；领域最近邻
中图分类号：TP311 **文献标识码**：A **文章编号**：1007－9807(2012)05－0055－10

0 引 言

在竞争日趋激烈的当今社会，销售环境的选择举足轻重，良好的销售环境有助于降低产品价格和厂商的生产规模，促进产品多样性[1]，互联网无疑就是这样良好的销售环境. 它的迅速普及促成了电子商务的日趋盛行，但它为消费者带来更多商品选择的同时，信息迷航的问题也日趋严重. 推荐系统的出现和应用极大地改善了这一问题. 一个准确、高效的推荐系统可以根据潜在消费者的个性化特征和需求，主动向其推荐商品，帮助他们找到所需商品，实现交叉销售[2]，顺利完成购买过程，增加顾客满意度的同时，获得和维持顾客忠诚度，最终促成重复购买[3]. 因此，网络推荐已成为电子商务网站成功实施网络营销的关键工具.

推荐系统包括基于内容过滤（content-based filtering，CBF）和协同过滤（collaborative filtering，CF）两大最为经典的推荐技术. 基于内容过滤的推荐结论直观且容易解释，不存在数据稀疏问题；协同过滤推荐则能够处理像音乐和视频等复杂的非文本结构化商品，发现用户的新兴趣[4]，推荐效果随着用户数量的增加而提高. 然而，基于内容推荐受商品属性提取方法的限制（例如不易对音乐和视频信息进行提取），存在新用户问题. 协同过滤推荐则主要存在着数据稀疏问题（sparsity problem）和扩展性[5]两大缺陷. 针对这些问题，一些研究结合两种算法提出了混合推荐算法. Albadvi 和 Shahbaze[6] 通过分析用户在各产品类别中的偏好，建立用户偏好模型，考虑用户与用户之间及用户与产品项之间的综合相似性形成最终推荐；张锋和常会友[7]采用 BP 神经网络预测用户对商品项的评分，减小候选最近邻数据集的稀疏性，极大的提高了预测的准确度，从而显著的提高了推荐质量；Ghazanfar 和 Prugel-Bennett[8] 利用用户对商品项的评分，商品的特征属性以及统计学信息进行混合推荐；Ariyoshi 和 Kamahara[9] 采用基于内容及基于协同过滤的双重奇异值分解法来缓解数据稀疏性问题；而王瑞琴和孔繁胜[10]则通过引入多个数据源对评价矩阵进行平滑填充来解决数据的稀疏性问题，并从用户和项目两方面进行联合聚类来

① 收稿日期：2011－08－16；修订日期：2012－02－15.
基金项目：国家自然科学基金资助项目(70801026)；教育部人文社会科学基金资助项目(07JA630054)；教育部博士点基金新教师资助项目(200805321007).
作者简介：朱国玮(1978—)，男，湖南长沙人，博士，副教授. Email: gwzhu@163.com

提高系统的可扩展性和推荐精度. 以上研究都在不同程度上改善了最终推荐效果，但没有一个能够同时克服内容过滤和协同过滤的各种缺陷，尤其在用户兴趣建模阶段，都忽略了用户兴趣随时间漂移问题. 对此，郑先荣等[11-12]在协同过滤算法中引入遗忘函数，通过不用的时间权重衡量用户兴趣的变化; 邢春晓等[13]在基于资源的协同过滤算法中引入基于时间和基于资源相似度的综合数据权重，提出适应用户兴趣变化的协同过滤推荐算法. 本文以这些研究为基础，提出一种新的混合推荐算法，从以下两个方面进行了改进及创新: 首先，采用基于内容的方法，考虑用户兴趣随时间变化，基于非线性逐步遗忘函数建立用户兴趣模型; 然后，将用户评分并集中的非目标用户划分为有推荐能力和无推荐能力两部分，省去无推荐能力用户与目标用户之间的相似性计算，对有推荐能力用户则采用领域最近邻协同过滤推荐算法，计算用户之间的综合相似性，预测目标用户未评价项的评分并做出推荐. 从实验结果来看，由于考虑了用户兴趣漂移问题，并进行了领域最近邻处理，新算法的推荐质量有了明显提高.

1 相关工作

1.1 用户兴趣模型的建立

在电子商务网站中，虽然用户和产品属性是隐含的，用户对商品的选择却并非是随机的[14]. 因此，无论是哪种推荐系统，它们都必须建立用户兴趣模型，才能据此做出推荐. 用户兴趣建模是获取和维护反映用户兴趣和知识的过程，其结果是建立能够表示用户特有知识背景、兴趣或需求的用户兴趣模型. 从建模时用户的参与程度来看，用户兴趣模型可分为显式建模、隐式建模以及显式和隐式混合建模. 显式用户兴趣建模由用户通过输入相关信息显式地建立，而隐式用户兴趣模型则是利用用户的隐式反馈信息来建立和改进的[15]. 然而，这些信息并不容易获取，用户往往不乐意付出这些额外的时间[16]，而通过观察用户的行为进行隐式推测，则可以减少用户不必要的负担[17-18].

用户兴趣建模主要有两个步骤: 首先，观察用户的行为，从中识别用户的兴趣. 然后，以此为基础，确定推荐系统下一步的反应. 当用户兴趣固定不变时，根据以上步骤就可以完成用户兴趣建模. 但是，当用户兴趣变化时，系统无法准确识别用户目标，严重影响推荐质量. 就现实而言，用户的兴趣不仅多样，且会随时间动态变化，跟踪和学习用户兴趣虽艰难但十分必要[19]. 因此在用户兴趣建模阶段，考虑用户兴趣与时间的关联关系，有利于实现更为准确的预测，从而实现有效推荐.

1.2 最近邻的寻找

传统的协同过滤推荐基于用户 - 商品评分矩阵 $\boldsymbol{R}(m,n)$ 为目标用户找寻最近邻. $\boldsymbol{R}(m,n)$ 表示 m 个用户对 n 个商品项的评分矩阵. 在大型电子商务网站中，用户及商品数目极其庞大，并随着网站的发展数目还在不断地增加，而用户评价过的商品项又很少，通常在 1% 以下[20]，这就使得 $\boldsymbol{R}(m,n)$ 成为高维稀疏矩阵，严重影响最终的推荐效果.

针对这一问题，有研究利用目标用户 u 和非目标用户 v 的评分项并集 I'_{uv} 计算两者之间的相似性[21]. 对于 u,v 在并集中未评价项 i，通过为各自寻找最近邻，利用这些最近邻的商品评分，预测 u,v 未评价项的评分，使得两者对并集中的所有项均有评分，再在此基础上进行进一步的协同过滤，并做出相应的商品推荐. 该方法有效地解决了评分矩阵的高稀疏性所导致的系列问题，提高了推荐质量. 但可以看到，该方法在寻找最近邻时，并未考虑到两个用户多样化的兴趣之间，有时并不存在绝对的相关性. 这是由于，在现实生活中，用户对于商品的效用评价，除商品本身的性能相关外，与其人际关系网络结构存在着更为紧密的关联[22]，两个用户在某个领域兴趣偏好相同，在其他领域，兴趣则未必也会相同. 因此，基于全体项目空间寻找最近邻的做法并不合适，可取的做法是: 基于未评价项所属类别为目标用户找寻同领域内的最近邻，以这些最近邻的商品评分作为进一步协同过滤的基础，此即“领域最近邻”处理方法[23]，通过该方法找到的最近邻更能代表目标用户的相似兴趣，也更符合实际生活情况.

进行了以上的思考和分析，分别在推荐模型的相应阶段，引入非线性逐步遗忘函数和领域最近邻处理方法，从而形成新的混合推荐模型，以期获得更佳的推荐效果，实验结果为这一期待给予了充分的支撑.

2　基于遗忘函数的用户兴趣建模

基于用户兴趣与商品属性的密切相关性，将用户兴趣从简单的对商品的整体印象细化到对商品各属性的不同程度兴趣度，即对于商品的某种属性，用户的偏好除了绝对的肯定(值为 1) 及否定(值为 0) 外，其偏好程度还可表达为 0、1 之间的某个值(如 0.73)，该值计算公式如下

$$d(u_i, p_j) = \frac{\sum r(p_j)/m}{\sum r(u_i)/n} \tag{1}$$

其中，$d(u_i, p_j)$ 表示用户 u_i 对商品属性 p_j 的偏好程度，取值于 0 与 1 之间，函数值越接近 1 表示该用户对于具备属性 p_j 的商品的偏好程度越高; $\sum r(p_j)$ 表示用户 u_i 评价过的具有属性 p_j 的商品评分之和; $\sum r(u_i)$ 表示用户 u_i 全部评分之和; m、n 分别为相应的商品评分项数.

用户的消费过程实际上是个动态决策的过程，在这个决策过程中，用户除了考虑当期的效用外，还会同时考虑当期决策对于未来决策的影响[24]. 就是说，用户的消费行为和购买决策在时间上是相互关联的，用户近期访问的商品更能影响用户当前的购买决策，也更能反映用户当前的兴趣，早期访问的商品对于用户未来可能产生兴趣的商品的影响作用较小. 而根据德国心理学家艾宾浩斯(Ebbinghaus) 对遗忘现象所做的系统研究，人对于事物的遗忘过程是非线性的(图 1)[25].

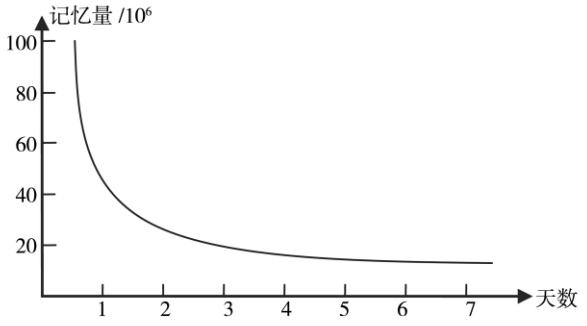

图 1　艾宾浩斯记忆遗忘曲线

Fig. 1 The Ebbinghaus forgetting curve

用户的消费购买行为是用户心理行为的反映，可以猜想用户对商品的兴趣变化也会遵循这样的遗忘规律. 因此，结合用户的兴趣随时间先快后慢变化的特征，在用户的商品属性偏好计算中引入非线性遗忘函数[12]

$$h(t_i) = (1-\theta) + \theta\left(\frac{t_i - t_{\min}}{t_{\max} - t_{\min}}\right)^2,$$

$$t_{\min} < t_i < t_{\max}, 0 \leqslant \theta \leqslant 1, 0 < h(t_i) < 1 \tag{2}$$

式中，t_i 为用户对第 i 个商品的访问时间与有效起始时间的间隔; $t_{\min}$ 为用户历史访问记录中的最早时间与有效起始时间的间隔; $t_{\max}$ 为用户历史访问记录中的最晚时间与有效起始时间的间隔; θ 为遗忘系数，其值反映了遗忘速度，即 θ 值越大，遗忘速度越快，用户的兴趣变化也就越快. 当 $\theta = 0$，对用户的历史评分未进行非线性遗忘处理; 当 $0 < \theta < 1$ 时，进行部分的非线性遗忘处理; 当 $\theta = 1$，对用户的历史评分进行完全的非线性遗忘处理. θ 值的设定应结合推荐系统中的用户兴趣变化的速度，兴趣变化快，θ 值应设得大一些，慢则应小些. 于是，用户对已评价商品的各属性偏好值调整为

$$D(u_i, p_{q,j}) = b_j d(u_i, p_j)\, h(t_i) \tag{3}$$

式中，当某商品 q 具备某个属性 j 时，b_j 取值为 1，否则该值为 0. 结合用户的商品评分历史记录及计算出的用户对商品的各属性偏好值，建立如下回归模型

$$r_{u_i,q} = \omega_{i,0} + \omega_{i,1} D(u_i, p_{q,1}) + \omega_{i,2} D(u_i, p_{q,2}) + \cdots + \omega_{i,j} D(u_i, p_{q,j}) + \cdots + \omega_{i,k} D(u_i, p_{q,k}) \tag{4}$$

在该模型中，$r_{u_i,q}$ 是用户 u_i 对商品 q 的总评分; k 是该商品的属性个数. 通过该模型获得的各参数 $\omega_{i,x}(x \in [0,k])$，即为各属性的权重，反映了商品各属性对用户兴趣的重要度.

3　基于领域最近邻的协同过滤推荐

3.1　领域最近邻的界定

通过基于内容的方法获得了“用户 - 商品

评分矩阵”之后，就需要为目标用户寻找最相近邻居，传统的处理方法是等同的计算目标用户与矩阵中每个用户之间的相似性，按照相似性的高低确定最近邻，预测目标用户未评价商品的评分，根据预测评分做出商品推荐. 然而，在实际的商务网站中，尤其是大型商务网站，由于商品数量的极其庞大，其商品都是进行类别划分的. 可以看到，多数用户的商品评分都随偏好集中于少数几个特定类别中. 访问过同类别商品的用户之间，相似性相对更高. 因此，将相似性计算缩小到少数几个商品类别，而非整个商品空间，并按照目标用户未评分商品所属类别，在同领域内为其寻找最相近邻居，将有效提高推荐效果和效率，也更符合生活实际，这就是领域最近邻的基本思想.

定义 设目标用户 u 和有推荐能力用户 v 的评分项所属商品类别集合分别为 C_u 和 C_v，两者的评分商品类别交集为 $C_1 = \{C_1, C_2, C_3, \cdots, C_f\}$，则对于 $\forall C_j \in C_1 (1 \leqslant i \leqslant f)$，从“用户 - 商品评分矩阵”中抽取所有该类别的评分项及对应的用户组成新的评分矩阵 R_j，计算目标用户 u 与矩阵中的其他用户之间的相似性，则其中相似性最大的前 N 位用户为目标用户的领域最近邻[23].

根据定义，要寻找目标用户的领域最近邻，只需分析目标用户 u 与用户 v 的评分集合 I_u 与 I_v，以及它们的评分项并集 I_{uv}（以 5 分的评分制为例）.

1) 若 $I_{uv} \subset I_u$，即用户 v 的所有评价项，用户 u 都已经评价，则 v 相对于 u 属于无推荐能力用户;

2) 若 $I_{uv} \not\subset I_u$，即用户 v 对 u 的未评价项进行了评分，则分为以下两种情况考虑:

① 若 $r(v) \leqslant 2.5$（评分制的中间值），视用户 v 对于商品 i 无偏好倾向，则 v 相对于 u 仍无推荐能力;

② 若 $r(v) > 2.5$，则表明用户 v 对商品 i 偏好明确，相对于 u 属于有推荐能力用户.

例如，为预测表 1 中目标用户 u 对商品类别 A 中商品项 I_3 的评分，需要在对商品项评价过的用户集合 $v_i (1 \leqslant i \leqslant q)$ 中寻找领域最近邻. 由于用户未对商品 I_3 评分，搜寻范围缩减为 $\{v_1, v_2, v_4\}$，而用户 v_2 对于商品 I_3 的评分低于 2.5，于是领域最近邻搜寻范围锁定为 $\{v_1, v_4\}$，极大的减少了相似性计算量.

表 1 用户 - 商品评分矩阵

Table 1 User-product rating matrix

用户	商品类别 A					
	I_1	I_2	I_3	I_4	I_5	I_6
u	4	3		1	2	5
v_1	1		5	5		4
v_2	1		2	5	3	
v_3	5	1		2	4	1
v_4	3	2	4	3	1	

3.2 相似性计算

按照领域最近邻处理方法，确定了目标用户的领域最近邻集合 $V_c = \{v_1, v_2, \cdots, v_q\}$ 后，利用“用户 - 商品评分偏好矩阵”以及人口统计信息计算目标用户与各领域最近邻的综合相似性.

以评分矩阵为基础，采用有约束的皮尔逊相关系数度量目标用户与领域最近邻集合中的各用户 $v_i (1 \leqslant i \leqslant q)$ 之间的相似性，公式如下

$$sim_1(u, v) = \frac{\sum_{i \in I_{u,v}} (r_{u,i} - m)(r_{v,i} - m)}{\sqrt{\sum_{i \in I_{u,v}} (r_{u,i} - m)^2 \sum_{i \in I_{u,v}} (r_{v,i} - m)^2}} \quad (5)$$

其中，m 表示评分制的中值（例如 5 分制的中值 2.5 分），I_{uv} 表示 u, v 的评价商品项交集.

目标用户与领域最近邻的人口统计信息之间的相似性计算如下

$$sim_2(u, v) = \sum_{i=1}^{n} \left(1 - \frac{|f_{u,i} - f_{v,i}|}{F_i}\right) \lambda_i \quad (6)$$

其中 $f_{u,i}$ 与 $f_{v,i}$ 分别表示用户 u 与 v 的第 i 个人口统计信息值，F_i 表示第 i 个人口统计信息的取值范围，λ_i 为第 i 个统计信息在相似性计算中的权值，n 为人口统计信息属性的个数.

最后，对以上两类相似性赋予一定的权值，而获得目标用户与各领域最近邻之间的综合相似性，计算公式如下

$$sim(u, v) = \beta sim_1(u, v) + (1 - \beta) sim_2(u, v) \quad (7)$$

其中，β 为可调节的权重因子. 当领域内的某用户

与目标用户之间的公共评价商品项数很多，仅依据评分矩阵就能获得非常准确的相似性，不需要引入人口统计信息以提高预测准确性，此时 β 取值为 1; 反之，若领域内的某用户与目标用户之间的公共评价商品项数有限，不足以真实反应用户之间的相似性，此时，应提高人口统计信息相似性的权重，甚至将 β 取值为 0 而直接以人口统计信息度量用户之间的相似性.

按照以上方法，计算出目标用户 u 与领域最近邻集合中的每位用户 v 之间的相似性 $sim(u,v)$，取相似性由大到小排列的前 k 位用户形成目标用户 u 的最近邻集合 $V=\{v_1,v_2,\cdots,v_k\}$.

3.3　推荐的生成

获得了最近邻集合后，设 V 中各用户的商品评分集分别为 $I_1,I_2,\cdots,I_q$，目标用户 u 的商品评分集为 I_u，并通过以下公式预测 u 未评价商品项的评分

$$R_{u,i}=\bar{r}_u+\frac{\sum_{v_q\in U} sim(u,v_q)\,(r_{v_q,i}-\bar{r}_{v_q})}{\sum_{v_q\in U}(|\,sim(u,v_q)\,|)} \tag{8}$$

其中，$sim(u,v_q)$ 表示最近邻集合中的用户 v_q 与目标用户 u 之间的相似性，$\bar{r}_u$ 表示 u 在所评价过的商品集中的平均评分，$r_{v_q,i}$ 表示 v_q 对商品 i 的评分，$\bar{r}_{v_q}$ 表示 v_q 在与 u 的公共评价商品集中的平均评分. 依此预测出目标用户在领域中所有未评价项的评分后，按 $R_{u,i}$ 值的大小取前 N 位形成 top－N 商品推荐集推荐给目标用户，完成整个推荐过程.

3.4　算法描述

本文所提出的基于非线性逐步遗忘和领域最近邻的混合推荐模型的输入数据包括: m 个用户对 n 个商品的评分矩阵 $\boldsymbol{R}(m,n)$，以及相应的评分时间，遗忘因子 θ，各用户的人口统计信息，相似性的权重因子 β，最近邻用户数 k，推荐集的规模 N. 输出的则是针对目标用户 u 的“top-N”推荐集. 整个推荐过程可以划分为 4 个阶段.

阶段 1　建立用户兴趣模型. 根据评分信息以及商品的特征属性，利用式(1) 计算各用户对商品各属性的相对兴趣，利用式(2) 计算用户对各评分项的遗忘函数，结合式(3)、(4) 建立多元回归方程，求解方程，形成相应的用户兴趣模型.

阶段 2　填充评分矩阵. 结合已建立的用户兴趣模型以及用户未评价商品的属性，预测用户对相应商品的评分，填充“用户 － 商品评分矩阵”，从而降低其稀疏性.

阶段 3　查找领域最近邻. 针对目标用户 u，判断用户集合的每个用户 v 是否为 u 的有推荐能力用户. 若 v 为 u 的有推荐能力用户，则按照领域最近邻的处理方法利用式(7) 计算两者的综合相似性，按照相似性高低选取前 k 位作为目标用户的领域最近邻.

阶段 4　预测评分与生成推荐. 利用目标用户和他的领域最近邻的综合相似性以及相应商品评分，通过式(8) 预测目标用户未评价商品的评分，根据预测所得评分高低，选取前 N 种商品形成“top-N”推荐清单推荐给目标用户.

4　实验结果与分析

4.1　实验数据

采用 Movielens(http: //grouplens. umn. edu) 网站提供的数据集对本文提出的改进算法和传统的基于内容的算法以及传统的协同过滤算法进行分析比较. Movielens 是美国明尼苏达州立大学 Grouplens 研究小组所开发的基于协同过滤算法的个性化电影推荐网站. Grouplens 研究小组公布了规模不同的 3 个数据集(http: //www. grouplens. org/data/)，所有电影分属于 19 种电影类别，每位用户至少对 20 部以上自己看过的电影进行评分，分值在 1 ~ 5 分之间，评分越高，意味着用户越偏爱该部电影. 本实验下载其中一个包含 943 位用户对 1 682 部电影的 100 000 条评分的数据集，从中随机抽取 100 个用户对 1 682 部电影的 11 560 条评分数据作为实验的数据基础，从实验数据集中的评分随机隐藏约 10%(1 155 条) 组成测试集，以其余约 90%(10 405 条) 的评分数据作为训练集. 根据实验需要，进一步在训练集中 3 次随机抽取 50 个用户，从而获得稀疏性不同的 4 个训练集，以此为基础预测各用户被隐藏电影的评分. 实验数据集的具体分布情况如表 2 以及图 2 所示.

表 2 实验数据集

Table 2 Experimental data set

	总样本	训练集 1	训练集 2	训练集 3	训练集 4	测试集
用户数	100	100	50	50	50	97
电影数	1 682	1 265	1 098	1 075	1 265	567
评分数	11 560	10 405	5 354	5 051	10 405	1 155
稀疏性	0.931 3	0.917 7	0.902 5	0.906 0	0.835 5	0.979 0

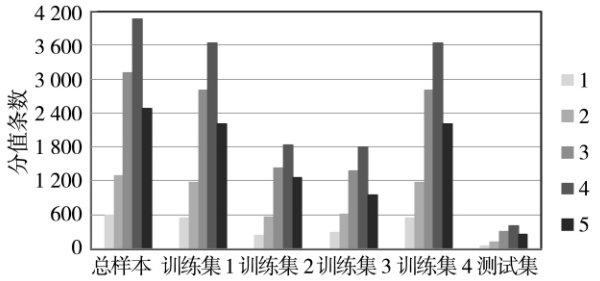

图 2 评分值分布

Fig. 2 Rating distribution

4.2 评价标准

推荐质量的评价标准主要有两类: 统计精度以及决策支持精度[20]. 本文采用广泛使用的平均绝对误差(MAE) 作为推荐效果的评价标准.

通过计算用户的预测评分与对应的实际评分之间的平均绝对偏差来度量模型预测的准确性,平均绝对误差越小,推荐质量越高. 假设用户的预测评分集为$\{p_1,p_2,\cdots,p_n\}$,对应的实际评分集为$\{r_1,r_2,\cdots,r_n\}$,MAE 的计算为

$$MAE = \frac{\sum_{k=1}^{n} | p_k - r_k |}{n} \tag{9}$$

4.3 实验结果及分析

为了测量新模型各阶段的关键因素对最终推荐效果的影响,进行了不同的对比实验,所得结果及分析如下.

1) 在建立用户兴趣模型阶段,基于遗忘函数对训练集中用户未评分项进行评分预测. 首先计算出训练集中各用户对各种电影属性的偏好值,以 $ID = 513$ 为例,数据如表 3 所示. 将遗忘因子 θ 分别设为 0(此时未进行非线性遗忘处理) 、0.3 和 0.6,遗忘因子越大,遗忘速度越快,针对每个用户求解而得 3 个不同的用户兴趣模型,并直接利用这些用户兴趣模型对测试组进行评分预测,以测量遗忘函数对推荐效果的影响.

表 3 $ID = 513$ 用户的电影评分及相应电影的属性分布

Table 3 Films ratings of user $ID = 513$'s and the corresponding film attribute distribution

用户 ID	电影项	评分	电影属性																	
			I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}
513	50	5	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0
513	117	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
513	118	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
513	121	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
513	127	4	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
513	181	5	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0
513	210	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
513	222	5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
513	250	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
513	323	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
513	405	3	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
513	435	5	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
513	472	4	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
513	546	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
513	685	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
513	841	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

表 4　*ID* = 513 用户的属性偏好值

Table 4 Attribute preference values of user *ID* = 513

*ID*513	I_1	I_2	I_5	I_6	I_8	I_9	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}
属性偏好值	1.01	1.04	0.99	0.91	0.91	0.91	0.68	1.14	1.04	1.03	1.14	1.14

注: 对该属性的偏好为 0,即它对具备该属性的电影不存在偏好倾向.

表 5　*ID* = 513 用户的遗忘函数值 $h(t_i)$ 的部分计算结果

Table 5 Parts of the calculated results of user *ID* = 513's forgetting function $h(t_i)$

	50	117	118	181	252	265	435	472	841
m = 0	0.957 7	0.957 7	0.957 7	0.957 7	0.966 1	0.957 7	0.966 1	0.957 7	0.957 7
m = 0.3	0.881 4	0.883 7	0.884 7	0.881 2	0.966 1	0.900 4	0.939 0	0.887 0	0.885 9
m = 0.6	0.912 0	0.913 3	0.913 9	0.911 8	0.966 1	0.923 3	0.949 9	0.915 3	0.914 7

表 6　训练集 1 未进行矩阵填充的部分评分矩阵

Table 6 Parts of the rating matrix of training set 1 before filling

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	4	0	0	4	0	4	0	4	5	3	0
20	3	0	0	0	0	0	0	0	0	0	2	0	0	0
23	5	0	0	0	0	0	4	0	0	0	0	0	4	4
25	5	0	0	0	0	0	4	4	0	0	0	0	0	0

表 7　基于遗忘函数建立回归模型进行矩阵填充后训练集 1 的部分评分矩阵

Table 7 Parts of the rating matrix of training set 1 after filling through regression modeling based on forgetting function

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
3	2.28	3.89	1.81	3.62	2.93	3.10	2.71	3.24	3.10	2.61	1.97	1.97	2.28	3.79
10	4.01	4.18	4.03	4	4.42	4.22	4	4.66	4	4.53	4	5	3	4.25
20	3	3.85	1.58	6.34	3.82	4.01	3.77	5.13	4.01	2.76	2	1.97	2.41	3.95
23	5	3.77	3.70	3.20	4.30	3.67	4	3.25	3.67	3.63	4.08	4.08	4	4
25	5	4.08	4.34	4.09	3.60	3.93	4	4	3.93	3.57	3.84	3.84	4.50	4.08

实验结果表明,加入了遗忘函数的用户兴趣模型的预测准确性要优于未加入遗忘函数的用户兴趣模型,如图 3 所示. 就总体而言,当遗忘因子取得较大时,推荐效果有所改善,原因可能是多数用户的兴趣变化较快,而个别用户(如用户 *ID* = 558) 可能由于兴趣变化慢,在加入遗忘函数后,推荐效果反而变差.

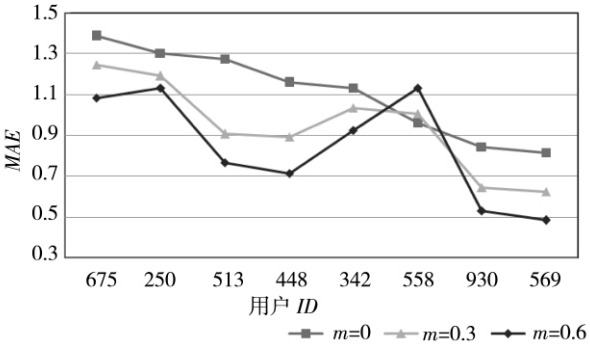

图 3　不同遗忘因子下的推荐精度比较

Fig. 3 Recommendation accuracy comparisons based on different values of forgetting factor

2) **鉴于以上结果,采用遗忘因子 θ 为 0.6 求解而得的各用户兴趣模型,预测训练集中各用户未评价电影的评分,以降低"用户 – 电影评分矩阵"的稀疏性. 在最近邻查找阶段,同时测量了未进行领域最近邻处理的基于内容过滤方法和本文进行了领域最近邻处理方法的推荐精度以作对比. 在计算综合相似性时,又将 β 值分别设为 0.1, 0.5 以及 0.9,以观测不同的相似性测度标准对于推荐精度的影响,结果显示如图 4,其中上半部分 3 条折线对应的是未进行领域最近邻处理的基于**

内容过滤的推荐精度，下半部分为进行了领域最近邻处理的本文方法的推荐精度. 由图 4 可知，进行了领域最近邻处理的本文算法具有更小的 *MAE* 值，这是由于用户可能存在多个不相关的兴趣领域，采用领域最近邻方法找到的最近邻与目标用户的兴趣更为相近，使得预测结果更为准确，从而提高了推荐质量. 而 β 取值为 0.9 时 *MAE* 值最小，则是由于本文第 1 阶段利用用户兴趣模型预测用户未评价电影的评分，极大的降低了评分矩阵的稀疏性，因此依赖评分矩阵进行预测使得预测结果更为准确.

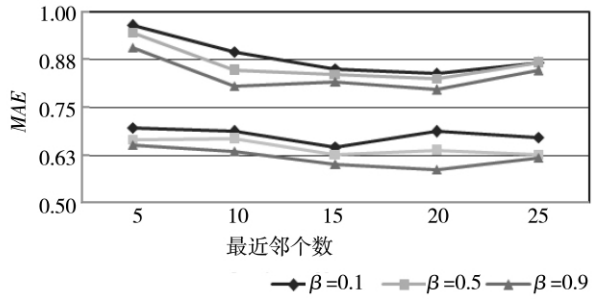

图 4　本文方法与未进行领域最近邻处理的 CBF 的推荐精度比较

Fig. 4 Recommendation accuracy comparison between proposed method and CBF without processing of domain nearest neighbor

3) 此外，还测量了本文的新模型与未进行基于遗忘函数的矩阵填充处理的协同过滤方法的推荐精度，对比结果如图 5 所示，其中，上半部分为传统协同过滤方法的推荐精度. 由图可以看出，相较于未进行基于遗忘函数的矩阵填充处理的协同过滤方法，本文的新模型在推荐精度有了大幅度的提高. 这是由于本文所提出的新模型综合了基于内容过滤的优势，利用基于非线性遗忘用户兴趣模型对用户未评分电影项进行预测，降低了评分矩阵的稀疏性之后，在为目标用户寻找最近邻阶段又引入了领域最近邻的处理方法，更进一步提高了推荐精度，另外利用人口统计信息的相似性、以及评分相似性获得综合相似性，都有利于改善预测准确性.

4) 最后，测量了在 3 组稀疏性不同的评分矩阵(相应的稀疏性为 0.906、0.902 5、0.835 5)，分别为各目标用户取最相邻的 5、10、15、20 以及 25 个领域最近邻情况下推荐精度的差异，实验结果见图 6. 可以看到，随着稀疏性的下降，*MAE* 值也在减小，即推荐的准确性在增强. 并且当领域最近邻个数选取为 20 个时，*MAE* 值达到最小，结果如图 6 所示.

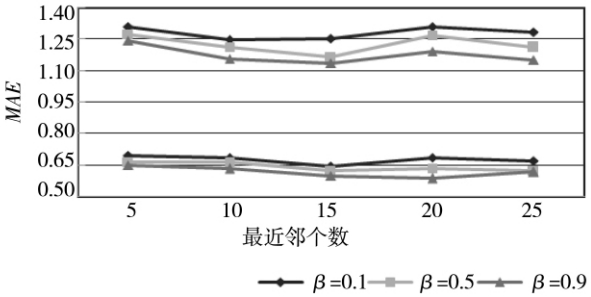

图 5　本文方法与未进行基于遗忘函数的矩阵填充的 CF 的推荐精度比较

Fig. 5 Recommendation accuracy comparison between proposed method and CF without matrix filling based on forgetting function

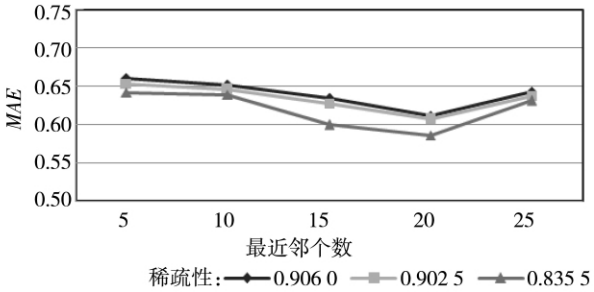

图 6　不同邻居个数下的推荐精度比较

Fig. 6 Recommendation accuracy comparisons based on different neighbor numbers

5　结束语

针对现有推荐系统存在的一些缺陷（如基于内容过滤的新用户问题，协同过滤的数据稀疏性、冷启动问题），以及用户兴趣变化的不确定性，本文提出了一套解决方案. 首先，引入非线性遗忘函数建立用户兴趣模型，采用基于内容的方法预测用户未评价项的评分，降低评分矩阵的稀疏性. 在最近邻查找阶段，考虑用户兴趣的多样性与非相关性，引入领域最近邻处理方法，并综合利用人口统计信息以及评分矩阵计算目标用户与各领域最近邻间的综合相似性，预测目标用户的未评价项商品评分，依据评分做出相应推荐. 可以看出，本文方法利用到了用户的人口统计信息以及商品的项目属性信息，并非仅仅只是简单依赖评分矩阵做出推荐，同时解决了新用户

和冷启动问题. 从实验结果来看，本文方法较传统协同过滤方法有着更高的预测推荐精度. 进一步的研究工作将在用户兴趣模型建立阶段将引入更能表征用户兴趣变化的方法，以建立更能准确的反应用户的当期兴趣的用户兴趣模型，提高预测推荐精确性.

参 考 文 献:

[1] 周　浩，朱卫平. 销售成本、垄断竞争与产品多样性 [J]. 管理科学学报，2008，11(5)：23－32.
Zhou Hao，Zhu Weiping. Sales cost，monopolistic competition and product diversity [J]. Journal of Management Sciences in China，2008，11(5)：23－32. (in Chinese)

[2] Fleder D，Hosanagar K. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity [J]. Management Science，2009，55(5)：697－712.

[3] 黄劲松，赵　平，王　高等. 中国顾客重复购买意向的多水平研究 [J]. 管理科学学报，2004，7(6)：79－86.
Huang Jinsong，Zhao Ping，Wang Gao，et al. Study on customers' repurchase intention in China [J]. Journal of Management Sciences in China，2004，7(6)：79－86. (in Chinese)

[4] Jung J H，Schneider C，Valacich J. Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments [J]. Management Science，2010，56(4)：724－742.

[5] Schafer J B，Konstan J，Riedi J. Recommender systems in E-commerce [C] //Proceedings of the 1st ACM Conference on Electronic Commerce，New York: ACM，1999: 158－166.

[6] Albadvi A，Shahbaze M. A hybrid recommendation technique based on product category attributes [J]. Expert System with Applications，2009，36(9)：11480－11488.

[7] 张　锋，常会友. 使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题 [J]. 计算机研究与发展，2006，43(4)：667－672.
Zhang Feng，Chang Huiyou. Employing BP neural networks to alleviate the sparsity issue in collaborative filtering recommendation algorithms [J]. Journal of Computer Research and Development，2006，43(4)：667－672. (in Chinese)

[8] Ghazanfar M A，Prugel-Bennett A. A scalable，accurate hybrid recommender system [C] //Third International Conference on Kowledge Discovery and Data Mining，Phuket: IEEE，2010: 94－98.

[9] Ariyoshi Y，Kamahara J. A hybrid recommendation method with double SVD reduction [J]. Database System For Advanced Applications，2010，(6193)：365－373.

[10] 王瑞琴，孔繁胜. 基于多数据源和联合聚类的智能推荐 [J]. 模式识别与人工智能，2008，21(6)：775－781.
Wang Ruiqin，Kong Fansheng. Intelligent recommendation based on multiple data and co-clustering [J]. Pattern Recognition and Tificial Intelligence，2008，21(6)：775－781. (in Chinese)

[11] 郑先荣，曹先彬. 线性逐步遗忘协同过滤算法的研究 [J]. 计算机工程，2007，33(6)：72－73，82.
Zheng Xianrong，Cao Xianbin. Research on lineal gradual forgetting collaborative filtering algorithm [J]. Computer Engineering，2007，33(6)：72－73，82. (in Chinese)

[12] 郑先荣，汤泽滢，曹先彬. 适应用户兴趣变化的非线性逐步遗忘协同过滤算法 [J]. 计算机辅助工程，2007，16(2)：69－73.
Zheng Xianrong，Tang Zeying，Cao Xianbin. Non-lineal gradual forgetting collaborative filtering algorithm capable of adapting to users' drifting interest [J]. Computer Aided Engineering，2007，16(2)：69－73. (in Chinese)

[13] 邢春晓，高凤荣，战思南 等. 适应用户兴趣变化的协同过滤推荐算法 [J]. 计算机研究与发展，2007，44(2)：296－301.
Xing Chunxiao，Gao Fengrong，Zhan Sinan，et al. A collaborative filtering recommendation algorithm incorporated with user interest change [J]. Journal of Computer Research and Development，2007，44(2)：296－301. (in Chinese)

[14] Huang Z，Zeng D D，Chen H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems [J]. Management Science，2007，53(7)：1146－1164.

[15] Moon S，Russell G J. Predicting product purchase from inferred customer similarity: An autologistic model approach [J].

Management Science, 2008, 54(1): 71 – 82.
[16] Carroll J, Rosson M B. Interfacing Thought: Cognitive Aspects of Human-Computer Interaction [M]. Cambridge, MA: MIT Press, 1987: 80 – 111.
[17] Dumais S, Joachims T, Bharat K, et al. Implicit measures of user interests and preferences [J]. ACM SIGIR Forum, 2003, 37(2): 50 – 54.
[18] Kelly D, Teevan J. Implicit feedback for inferring user preference: A bibliography [J]. ACM SIGIR Forum, 2003, 37(2): 18 – 28.
[19] 曾 春, 邢春晓, 周立柱. 个性化服务技术综述 [J]. 软件学报, 2002, 13(10): 1952 – 1961.
Zeng Chun, Xing Chunxiao, Zhou Lizhu. A review of personalized service technology [J]. Journal of Software, 2002, 13(10): 1952 – 1961. (in Chinese)
[20] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms [C] // Proceedings of the 10th International Conference on World Wide Web, New York: ACM, 2001: 285 – 295.
[21] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法 [J]. 软件学报, 2003, 14(9): 1621 – 1628.
Deng Ailin, Zhu Yangyong, Shi Baile. A collaborative filtering recommendation algorithm based on item rating prediction [J]. Journal of Software, 2003, 14(9): 1621 – 1628. (in Chinese)
[22] 黄玮强, 庄新田, 姚 爽. 网络外部性条件下新产品扩散的赠样策略研究 [J]. 管理科学学报, 2009, 12(4): 51 – 63.
Huang Weiqiang, Zhuang Xintian, Yao Shuang. Sampling strategies for diffusion of new products with network externality [J]. Journal of Management Sciences in China, 2009, 12(4): 51 – 63. (in Chinese)
[23] 李 聪, 梁昌勇, 马 丽. 基于领域最近邻的协同过滤推荐算法 [J]. 计算机研究与发展, 2008, 45(9): 1532 – 1538.
Li Cong, Liang Changyong, Ma Li. A collaborative filtering recommendation algorithm based on domain nearest neighbor [J]. Journal of Computer Research and Development, 2008, 45(9): 1532 – 1538. (in Chinese)
[24] 李纯青, 徐寅峰. 动态消费者选择模型及贴现因子的确定 [J]. 管理科学学报, 2005, 8(3): 50 – 55.
Li Chunqing, Xu Yinfeng. Determining of consumer choice models and discount factor [J]. Journal of Management Sciences in China, 2005, 8(3): 50 – 55. (in Chinese)
[25] Yuan Genqing. Medical Psychology [M]. Nanjing: Southeast University Press, 1995: 47 – 53.

Hybrid recommendation based on forgetting curve and domain nearest neighbor

ZHU Guo-wei, ZHOU Li
School of Business Administration, Hunan University, Changsha 410082, China

Abstract: Content-based filtering and collaborative filtering are the two most classical algorithms in recommendation system. However, there is the new customer problem in content filtering which does not consider the influences of users´ interests drifting on recommendation quality. And collaborative filtering faces severe challenges of data sparisity and cold start. To solve these problems, a hybrid recommendation algorithm is proposed in this paper. First, the paper builds a Customer Interests Model (CIM) based on the Forgetting Curve to predict the unevaluated rating; then introduces the processing method of "domain nearest neighbor" to find the nearest neighbors for target users to predict the unevaluated rating; and finally, makes the recommendation. The experimental results show that the proposed method can improve the recommendation quality effectively.
Key words: hybrid recommendation; non-lineal gradual forgetting; domain nearest neighbor