文章编号：1007-2373 (2010) 03-0082-06

一种改进的协同过滤推荐算法

杨 芳[1]，潘一飞[2]，李 杰[1]，王云峰[1]

（1. 河北工业大学 管理学院，天津 300401；2. 四川大学 电气信息学院，四川 成都 610065）

摘要 电子商务的蓬勃发展，使网站中能够提供的商品种类日益繁多，如何迎合客户的兴趣来推荐商品，成为当前电子商务亟待解决的重点问题. 协同过滤作为目前推荐系统应用中最为成功的个性化推荐技术，也得到了越来越多研究者的关注. 文章在简要介绍传统协同过滤推荐算法的基础上，重点对推荐算法无法适用于用户多兴趣下的推荐问题进行了剖析，提出了一种基于用户多兴趣的协同过滤推荐改进算法. 通过实验仿真，验证了该算法的有效性.

关 键 词 电子商务；个性化推荐；数据挖掘；协同过滤推荐算法；用户多兴趣

中图分类号 TP311.13 **文献标识码** A

An Adaptive Algorithm for Collaborative Filtering Recommendation

YANG Fang[1]，PAN Yi-fei[2]，LI Jie[1]，WANG Yun-feng[1]

（1. School of Management, Hebei University of Technology, Tianjin 300401, China; 2. School of Electrical Engineering and Information, Sichuan University, Sichuan Chengdu 610065, China）

Abstract The vigorous development of e-commerce has enabled websites to provide an increasingly larger variety of products. How to recommend products catering for the interest of customers has become an important issue that urgently requires today's e-commerce to solve. Collaborative filtering, as the most successful personalized technique of recommendation in the existing recommendation system application, has gained more and more researchers' attention. Based on a brief introduction of traditional collaborative filtering recommendation, the paper is focused on analysis of the problem that recommendation algorithm fails to apply to recommendation with the users' multiple interests, and proposes an improved algorithm of collaborative filtering based on multiple interests of users. Via experimental simulation, the validity of the algorithm is verified.

Key words electronic commerce; personalized recommendation; data mining; collaborative filtering recommendation algorithm; user multiple-interests

电子商务迅猛发展的时代已经到来，网上购物的交易方式已经彻底改变了传统的商业模式. 2008年，电子商务的交易额已达到了近两万亿人民币. 在网站这种虚拟的商务环境下商家所能提供的商品种类日益繁多，以国内知名的淘宝网为例，它跨越了 B2C 和 C2C 两种业务模式，所提供的商品从实体的书籍、衣物、家具到虚拟的游戏币、在线服务，种类可达 2 亿商品. 面对如此众多的商品，无疑会大大增加用户发现满意商品的困难，需要电子商务系统不断开发商品推荐功能来满足用户的需要.

1 传统的协同过滤推荐算法

协同过滤推荐算法是目前电子商务推荐系统中应用最为广泛和最为成功的个性化推荐技术. 传统的协同过滤算法基本思想是根据具有类似观点的用户的行为来对用户进行推荐或者预测.

协同过滤算法采用一个$M\times N$阶用户-项目评分矩阵R来表示用户输入的评分数据，使用统计技术寻

收稿日期：2009-09-25

基金项目：河北省自然科学基金（F2008000117）；河北省科技攻关项目（07213508D）

作者简介：杨芳（1981-），女（汉族），博士生.

找与目标用户有相同兴趣偏好的邻居，然后根据目标用户的邻居的兴趣偏好产生对目标用户的推荐．协同过滤算法主要有 3 个步骤：评分表示、邻居形成和推荐生成．

1.1 评分表示

在一个采用传统的协同过滤技术的推荐系统中，用户输入的评分数据可以用一个$M\times N$阶矩阵$R(M, N)$来表示，其中M行代表M个用户，N列代表N个项目，第i行第j列的元素$R_{i,j}$代表用户i对项目j的评分．评分可以使用 2 进制的 0 和 1 来表示用户的偏好（喜欢/不喜欢）或购买状态（已购买/未购买），也可以用数字分级表示用户对项目的喜好值．例如：MovieLens 中用户对电影的评分以从 0 到 5 之间的整数来表示用户的喜好．0 表示没有评分，1 到 5 数值越大表示用户的偏爱程度越大．

1.2 邻居形成

协同过滤算法的核心是根据用户-项目评分矩阵发现需要推荐服务的目标用户的最近邻，即：对当前用户u，要产生一个依照用户相似度大小进行排列的邻居的集合$N=\{N_1, N_2, \cdots, N_k\}$，$u$不属于$N$.

用户之间相似性的度量方法有许多种，最常用的两种方法是：相关相似性和余弦相似性．

相关相似性，可以通过 Pearson 相关系数公式来计算，主要用于衡量两个变量之间的线性关系；余弦相似性，用户评分可看作n维项目空间上的向量，用户间的相似性通过向量之间的余弦夹角度量，余弦值越大表示两用户间相似程度越高．

1.3 推荐产生

基于目标用户的最近邻，可以计算两类推荐结果包括用户对任意项的预测值和 Top-N 推荐集．

用户对任意项的预测值：已知用户u已评分项集I_u，则用户u对任意未评价项$k'(k'not\in I_u)$的预测值为

$$P_{u,k'}=\overline{R}_u+\frac{\sum_{m=1}^{N} Sim(u,m)*(R_{m,k'}-\overline{R}_m)}{\sum_{m=1}^{N} Sim(u,m)} \tag{1}$$

其中：$\overline{R}_u$是用户u对已评价项目的平均评分；$Sim(u, m)$是用户u与最近邻集合N中的用户m的相似系数；$R_{m,k'}$是用户m对用户u未评价项k'的评分，$\overline{R}_m$是用户m对已评价项的平均评分；N是最近邻的个数．$P_{u,k'}$值越大表示用户u对未评价项目k'的可能偏爱程度越高．

Top-N 推荐集：分别计算出用户u对未评分项目k'的预测值$P_{u,k'}$之后，对$P_{u,k'}$按降序进行排序，把排名靠前N个项目，提供给目标用户，作为该用户 Top-N 推荐集．

协同过滤技术在实际应用中获得了极大的成功[1]，一些大型的电子商务系统，如全球最大的网上书店 Amazon.com，最大的网上音乐商店 CDNow.com 就是将协同过滤技术应用到电子商务领域的成功案例．然而随着电子商务系统规模的扩大，用户数目和项目数在呈指数级增长，传统的协同过滤推荐算法需要给予整个用户和项目空间查询目标用户的最近邻居，这对算法本身无疑是一个极大的挑战[2]．

2　改进的协同过滤推荐算法

2.1 问题分析

传统的协同过滤推荐算法是用邻居用户对某一项目的偏好信息来判断用户对该项目的偏好，邻居用户是和当前用户具有相似兴趣爱好的用户．但在传统的协同过滤推荐算法中，邻居用户和当前用户的共同兴趣爱好并不一定是要预测的项目方面的兴趣爱好，而可能是其他方面的兴趣爱好．如果仍然用这些邻居用户来预测，其误差可想而知．如表 1 所示，这里共有 7 个用户，6 个项目，其中I_1、I_3、I_6都是英语方面的项目，假定项目不同，但都是有关英语的项目，而I_2、I_4、I_5是足球方面的项目，现要预测$R_{7,6}$=?.

表 1　用户-项目评价表例 I [3,4]

Tab. 1　An example of user-item data matrixI

User/Item	*I*1	*I*2	*I*3	*I*4	*I*5	*I*6
	英语	足球	英语	足球	足球	英语
*U*1	3	1	2	3	5	5
*U*2	3	1	2	3	5	5
*U*3	3	1	2	3	5	5
*U*4	1	5	3	3	1	1
*U*5	2	5	2	3	2	1
*U*6	3	5	1	3	2	1
*U*7	3	5	2	4	2	?

根据传统协同过滤推荐算法，将根据前面6个用户与第7个用户在I_1～I_5方面的兴趣爱好的相似性来决定邻居用户．如果规定只有3个邻居用户，则U_4、U_5、U_6将成为U_7的邻居用户，从而对$R_{7,6}$=？的预测将根据这3个用户在I_6方面的兴趣来计算，则显然$R_{7,6}=1$．但值得关注的是，之所以用户U_4、U_5、U_6成为U_7的邻居用户，主要是因为它们在足球方面有着共同的兴趣偏好，但要预测的是用户U_7在英语方面的兴趣偏好，如果根据足球方面的偏好信息来预测英语方面的偏好，推荐结果将受到怀疑．

表2 用户-项目评价表例II [3-4]

Tab. 2 An example of user-item data matrixII

User/ Item	I1 英语	I2 足球	I3 英语	I4 足球	I5 足球	I6 英语
U1	3	1	2	3	5	5
U2	3	1	2	3	5	5
U3	3	1	2	3	5	5
U4	0	5	0	3	1	1
U5	0	5	0	3	2	1
U6	0	5	0	3	2	1
U7	3	5	2	4	2	?

再假设一种极端情况，用户U_4、U_5、U_6可能只对足球方面的信息感兴趣，在项目I_1、I_3上没有评分，如表2所示，这样根据传统协同过滤推荐算法，U_7的邻居用户是用户U_4、U_5、U_6，这样对$R_{7,6}$=？的预测完全是根据U_4、U_5、U_6三用户在项目I_2、I_4、I_5方面的兴趣得到的，也就是对英语项目的预测完全是根据足球项目方面的信息得到，由于两类项目的差异性，势必会大大降低预测的准确率．

2.2 提出依据

从以上的问题分析中，可知用户多兴趣是现实存在的，而传统的协同过滤推荐算法只适合用户单一兴趣下的推荐，同时也注意到，对大多数用户来说，他们的兴趣一般只集中在某几个领域，只对感兴趣的项目进行评价，所以用户对项目的评价数据往往会集中出现在某几个类别中．

由此提出基于用户多兴趣的协同过滤推荐改进算法的研究假设：

1）用户是可以按兴趣分类的，并且兴趣之间存在差异；

2）用户对项目的浏览、评价包含了用户的兴趣信息；

3）用户对未知项目的评价和相似用户的评价相似．

那么仍然对于表1的示例，如果根据用户U_1~U_6与U_7在英语方面有相似兴趣偏好的用户来预测I_6，预测结果将更具有说服力，即I_1、I_3来计算邻居用户，得到英语方面的邻居用户U_1、U_2、U_3，从而来预测$R_{7,6}$=5．可见，如果能把对同一类别项目具有兴趣的用户聚到一类中，在小矩阵中寻找目标用户的最近邻居，一方面可以提高准确率，另一方面也可以大大减少计算量，与此同时由于用户在兴趣领域内的评分往往比较集中，亦可以有效地降低数据的稀疏性．

基于此，对传统协同过滤算法作如下改进：1）对用户兴趣进行分类，在系统中用户的兴趣是通过对项目的选择来进行了解的，所以把对用户兴趣的分类转化为对项目的分类，引入用户兴趣度的概念，来探讨用户在不同类别项目中所表现出来的兴趣差异，进而可以实现对用户多兴趣的了解；2）对于同一用户，如预测项目所属类别不同，用来预测的邻居用户也不同，也就是邻居用户与待预测的项目在内容上具有一定相似性，从而保证用来预测的邻居用户与当前用户在待预测项目上具有相似的兴趣爱好；3）用户具有多兴趣性，但用户对每类项目的兴趣也是不尽相同的，在推荐集中考虑以用户对不同类别项目的兴趣度作为权重，来分配每类项目的推荐数目．

2.3 算法设计

首先将项目采用某种技术按照某种标准划分为不同类别，然后把对此类项目有评价的用户的评价信息映射到此类，统计参数，计算用户在每类项目的兴趣度，当超过阈值时，认为该用户对该类项目有兴趣偏好，并由这些用户形成聚类，从聚类中搜寻针对此类项目的邻居用户产生推荐．详细过程如下：

1）按照分类规则对项目进行分类 把整个项目空间划分成若干类别，每个项目可能属于多个类别，每个类别包含至少一个项目．目前对项目进行分类有众多方法．比如对于图书的分类可以通过知识组织框架来实现．熊馨等（2005）[5] 提出了概念分层的方法对项目进行分类．还可以利用分类和聚类技术自动生成项目类别．但在实际的推荐系统中，项目的描述信息普遍存在着对项目所属类别的描述．比如文中的实验数据集，项目有明确的分类体系，故直接将其作为了项目分类的依据．

2）映射评价信息，统计参数，计算用户兴趣度，建立用户兴趣度矩阵，构造用户兴趣偏好特征用户兴趣度$A_{i,j}$，即用户i对项目类别j的兴趣度，衡量用户对某一类别项目的兴趣偏好.

$$A_{i,j}=\frac{M_{i,j}\sum_{j=1}N_j}{N_j\sum_{j=1}M_{i,j}} \tag{2}$$

其中：$M_{i,j}$表示用户i所评价的项目类别j中的项目数目；N_k表示项目类别k中包含的项目数目.

当$A_{i,j}=0$时表示用户i尚未评价过j类项目，在此之前未对j类项目表现出兴趣. 当$0<A_{i,j}\leqslant 1$时表示用户i对j类项目所表现出的兴趣不大或很一般，系统可以对此进行忽略. 当$A_{i,j}>1$时表示用户i对j类项目表现出了明显的兴趣，并且$A_{i,j}$越大，表明用户i对j类项目的兴趣偏好越大.

公式 (2) 在计算中同时考虑了某一项目类别中包含的项目数量大小和某一用户所评价的项目数量大小对用户兴趣度产生的影响. 所以并不是某个用户评价的某类项目多，就一定代表着用户对该类项目的兴趣偏好就大，还要考虑该类项目所包含的项目总量大小.

然后计算出用户对所有类别的用户兴趣度，可以形成用户兴趣度矩阵.

3）依据用户兴趣度矩阵和项目类别体系，进行用户聚类，形成用户兴趣模型

对于每一个用户i，判断用户U对项目的兴趣偏好分散在哪几个项目类别中，用户i在每个类别中的项目评分组成用户的兴趣描述，记为P_i

$$P_i=\{R_{i,j}|j=1,\cdots,J\} \tag{3}$$

其中：$R_{i,j}$表示在项目类别j中被用户i评价过的〈项目，评价〉集合.

汇总对j类项目有兴趣偏好的用户聚类中用户的$R_{i,j}$形成项目评价矩阵，记为Q_j

$$Q_j=\{R_{i,j}|i=1,\cdots,I\} \tag{4}$$

4）在用户聚类中计算用户间的相似度，寻找最近邻居

在项目评价矩阵Q_j中，计算对j类项目有兴趣偏好的用户聚类中用户间的相似性，即项目类别j中判断哪些用户与目标用户u相似，使用相关相似性公式，计算项目类别j中用户i与用户u之间的兴趣相似度$w^j(u,i)$

$$w^j(u,i)=\frac{\sum_{k\in I_{u,i}}(r_{u,k}-\bar{r}_u)(r_{i,k}-\bar{r}_i)}{\sqrt{\sum_{k\in I_{u,i}}(r_{u,k}-\bar{r}_u)^2}\sqrt{\sum_{k\in I_{u,i}}(r_{i,k}-\bar{r}_i)^2}} \tag{5}$$

对于用户u，把所有的$w^j(u,i)$进行排序，将排名在前$m_{u,j}$名用户作为用户u的最近邻居.

5）在用户聚类中计算用户为评价项的预测值

对于目标用户u，计算它在j类项目用户聚类中对未评价项目k的预测值$p^j(u,k)$.

$$p^j(u,k)=\bar{r}_u+\frac{\sum_{i=1}^{n}w^j(u,i)\times(r_{i,k}-\bar{r}_i)}{\sum_{i=1}^{n}w^j(u,i)}$$

如果项目j属于多个类别，则取预测值大者作为最终预测值.

6）产生推荐结果

对每一类的$p^j(u,k)$值进行降序排列，如果用户u的兴趣分散在L个项目类别中，推荐的项目总数为N，按用户对各项目类别的兴趣度权重计算推荐数量$N_{u,j}$.

$$N_{u,j}=\frac{A_{u,j}}{\sum_{m=j_1}^{J}A_{u,m}}\times\frac{M_{u,j}}{\sum_{n=j_1}^{J}M_{u,n}}$$

3 协同过滤推荐改进算法的实验与分析

3.1 数据集

实验使用的数据集来自Minnesota大学GroupLens Research项目组收集的MovieLens数据集.MovieLens站点（http: //www.movielens.umn.edu）是一个基于Web的研究型推荐系统，于1997年建立，系统接收用户对电影的评分并提供相应的电影推荐列表．目前，该Web站点的用户已经超过70000人，用户评分的电影超过5000部，电影种类包括12种．评分是从1到5的整数，数值越高，表明用户对该电影的偏爱程度越高．

实验从数据库中抽取了一部分数据集：包含350个用户对1 567部电影的27 475条评分数据．评分数据集被转换成一个用户-项目评分矩阵．每个用户至少评价了20部电影，并且包含了用户的简单人口统计学信息和电影的分类信息．在实验中也考虑了数据集的稀疏等级，它被定义为用户-评分矩阵中没有被评分的条目所占的百分比．数据集的稀疏等级为：1-27 475 /（350*1 567）=0.966 9．

3.2 度量标准

评价推荐系统推荐质量的度量主要包括统计精度度量方法和决策支持精度度量方法两类[6]．统计精度度量方法中平均绝对偏差 MAE（ mean absolute error ）易于理解，可以直观地对推荐质量进行度量，是最常用的一种推荐质量度量方法．文中采用平均绝对偏差MAE作为度量标准，通过计算预测的用户评分与实际用户评分之间的偏差度量预测准确性，MAE值越小，推荐精度越高．

3.3 实验结果与分析

协同过滤推荐算法在推荐系统中实现的效果，通常也会受到数据集稀疏程度和最近邻居用户个数两个因素的影响．因此，在验证改进算法的有效性时，围绕这两个因素重点对实验结果进行了对比．

1）数据的稀疏性比较

文中原始数据集的稀疏等级为0.9669．依据提出的基于用户多兴趣的协同过滤推荐改进算法中对用户兴趣度的度量，用户对电影类别的兴趣度大于阈值1，归入到对该类别电影有兴趣偏好的用户聚类中的原则，对用户-项目矩阵进行重新划分，依兴趣度划分之后，数据集的平均稀疏等级为0.8828，比原数据集的稀疏性有了明显改善．

2）最近邻居集大小不同下的推荐效果比较

邻居个数的变化对预测的质量有很大的影响．实验计算了在不同训练集与测试集所占比例下，邻居集个数不同时的MAE值，来验证推荐的质量．并对相同邻居个数下，不同训练集与测试集比例下的MAE值进行了加总平均，测试结果如表3和图1所示．可以看出随着邻居集数量的增加，两种算法的平均绝对偏差 MAE 的值逐渐减小，并且变化值会不断趋于平缓．相比较而言改进算法受邻居集数量的影响稍大一些，但是改进算法的预测质量在整个区间内始终好于传统协同过滤推荐算法．

表3 最近邻居集大小不同所得的测试数据

Tab. 3 Test data under different nearest-neighbor sets

邻居集大小	*MAE*	
	传统算法	改进算法
10	1.007 1	0.956 9
20	0.965 0	0.888 4
30	0.936 7	0.855 9
40	0.919 9	0.853 4
50	0.903 5	0.836 3

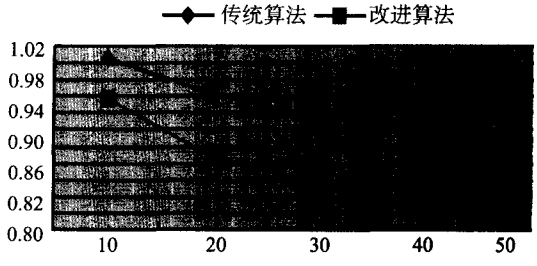

图1 最近邻居集大小不同所得的测试数据

Fig. 1 Test data under different nearest-neighbor sets

3）训练集与测试集不同比例下的推荐效果比较

实验计算了在不同的邻居集下，训练集与测试集比例不同时的*MAE*值，来验证推荐的质量．并对相同训练集与测试集比例下，不同邻居个数的*MAE*值进行了加总平均，测试结果如表4和图2所示．由图表可知，当训练集与测试集的比例不断增加的过程中，传统算法与改进算法的平均绝对偏差*MAE*的值都在减小，且改进算法的减少更为迅速，推荐质量提高明显．同时发现，随着训练集所占比例

的增大，*MAE* 的减少量在不断降低．但是改进算法的预测质量比传统算法在整个区间内都好一些．

通过对以上实验结果的分析，可以得出这样的结论：基于用户多兴趣的协同过滤推荐改进算法比传统的协同过滤推荐算法在所有的稀疏水平上都能提供更好的推荐质量．

表 4　训练集与测试集不同比例所得的测试数据

Tab. 4　Test data under train and test sets with different scales

训练集所占比例	*MAE*	
	传统算法	改进算法
0.2	0.9548	0.9438
0.4	0.9473	0.8840
0.6	0.9435	0.8502
0.8	0.9401	0.8347

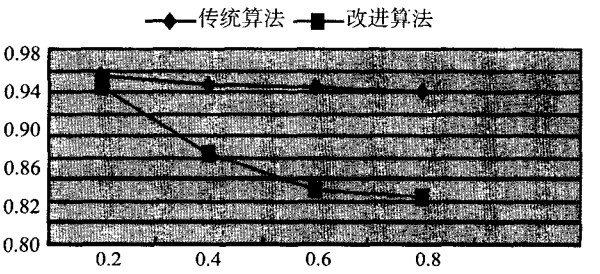

图 2　训练集与测试集不同比例所得的测试数据

Fig. 2　Test data under train and test sets with different scales

4　结语

电子商务推荐系统是个新兴的重点研究与应用领域．随着用户需求水平的提高，推荐算法与系统的研究在不断发展和完善[7]．本文提出的基于用户多兴趣的协同过滤推荐改进算法，正是为了解决现实中存在的用户多兴趣问题而产生的．通过与传统的协同过滤推荐算法对比，可知改进算法明显地改善了协同过滤推荐算法中常见的稀疏性问题，同时验证了改进算法比传统算法在所有的稀疏水平上都具有较高的推荐质量．

参考文献：

[1] Chen Y L，Cheng L C．A novel collaborative filtering approach for recommending ranked items [J]．Expert System with Applications，2008，34 (4)：2396-2405．

[2] Adomavicius G，Tuzhilin A．Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]．IEEE Trans on Knowledge and Data Engineering，2005，17 (6)：734-749．

[3] 余力，刘鲁，李雪峰．用户多兴趣下的个性化推荐算法研究 [J]．计算机集成制造系统，2004，10 (12)：1610-1615．

[4] 余力．电子商务个性化推荐若干问题的研究 [D]．北京：北京航空航天大学，2004．

[5] 熊馨，王卫平，叶跃祥．基于概念分层的个性化推荐算法 [J]．计算机应用，2005，25 (5)：1006-1009．

[6] Schafer J B，Konstan J A，Riedl J．Recommender systems in e-commerce [C]．In Proceedings of the First ACM Conference on Electronic Commerce [A]．Denver，CO：1999．158-166．

[7] 李杰，徐勇，王云峰，等．面向个性化推荐的强关联规则挖掘 [J]．系统工程理论与实践，2009，29 (8)：144-152．

[责任编辑　张颖志]