考虑时间和价格因素的 Web 客户需求协同推荐模型①

赵宏霞[1]，杨皎平[2]，万 君[1]

[1](辽宁工程技术大学 营销管理学院，葫芦岛 125105)
[2](渤海大学 管理学院，锦州 121013)

摘 要：协同过滤推荐算法是电子商务个性化推荐系统中采用最为广泛的推荐技术，但是传统的推荐方法在进行商品推荐时忽略了交易时间和产品的价格因素，从而导致推荐质量下降。针对这一问题，提出了考虑时间和价格因素的协同过滤模型，通过实验表明在计算 Pearson 相关系数时考虑时间和价格因素对算法的改进最为有效。

关键词：电子商务；推荐系统；协同过滤；时间；价格

Collaborative Filtering Recommendation Model of Web Customer Demand Considering Time and Price Factors

ZHAO Hong-Xia[1], YANG Jiao-Ping[2], WAN Jun[1]

[1](School of Marketing Management, Liaoning Technical University, Huludao 125105, China);
[2](College of Management, Bohai University, Jinzhou 121013, China)

Abstract: The collaborative filtering recommendation algorithm is the widely used technology in the personalized e-commerce recommendation system. However, the traditional recommendation algorithm neglected trading hours and product pricing when recommended products, which led to the lower quality recommended. To solve this problem, a collaborative filtering model considering time and price factors is proposed in this paper, and experiments show that the improvement of algorithm is most effective when time and price factors are taken into account in the calculation of Pearson correlation coefficient.

Key words: E-business; recommendation system; collaborative filtering; time; price

个性化推荐系统作为一个概念在 20 世纪 90 年代被提出[1,2]，随着 web2.0 技术的成熟，使得推荐系统得到了迅速的发展。通过该系统，电子商务网站能主动适应每个 web 客户的特定需求，为每个 web 客户提供了不尽相同的个性化购物环境，实现了电子商务中的“一对一营销”。目前，几乎所有大型的电子商务系统，如 eBay、Amazon、CDNOw、淘宝网、当当网等都不同程度地使用了各种形式的推荐系统。

基于最近邻居的协同过滤算法是目前电子商务推荐系统中运用最成功的算法之一，其基本思想是计算目标客户与 web 数据库中其余客户交易行为的相似性，搜索目标客户的最近邻居，然后用最近邻居的需求来预测目标客户的需求。

然而传统的协同过滤算法[3]在计算目标客户与基本客户相似性时，忽略了两个重要的因素：⑴没有考虑客户的交易时间。相关研究表明[4]，越是近期的交易评价参考价值越大；另外客户的兴趣偏好是随时间动态漂移的，要求推荐系统能实时的推荐新产品。②没有考虑交易价格的因素。高价产品交易后的评分比低价产品的评分更能表明客户的兴趣。针对第⑴个问题，王岚[6]和丛晓琪[7]进行了改进，对第二个问题尚未学者进行探讨。鉴于此，本文在传统协同过滤算法的基础上，提出了基于时间和价格加权的协同过滤算法，并利用实验数据集对传统算法和本文提出的改

① 基金项目:辽宁省教育厅项目(2009A326);中国煤炭工业协会项目(MTKJ2010-320),;教育部人文社科项目(10YJC630407)
收稿时间:2010-12-07;收到修改稿时间:2011-01-21

进算法在精确度上进行了比较。

1 相关工作

协同过滤算法基于这样的假设：若客户对一些项目(本文为商品)的评分比较相似，则他们对其它商品的评分也将会相似。

该算法实现需要满足的条件是：web 客户在交易后需要对此次交易的有关商品进行评分，进而 web 数据库中有 m 个客户对 n 个商品的评分矩阵 $R=(r_{ij})_{m\times n}$。

定义 1：推荐系统中的数据源 D=(U，I，R)，其中 $U=\{User_1, User_2, \cdots, User_m\}$是基本客户的集合，|U|=m; $I=\{Item_1, Item_2, \cdots, Item_n\}$是商品集合，|I|=n; m×n 阶矩阵 R 是基本客户对各商品的评分矩阵，其中的元素 r_{ij} 表示 U 中第 i 个客户对 I 中第 j 个商品的评分。

若目标客户 g 此前在该网站购买过 s(s<n)件商品，并且进行了评分。然后需要预测客户 g 对其余 n-s 件商品的评分，进而将评分预测值较高的商品推荐给客户 g。

预测客户 g 的评分，则需要搜索客户 g 的邻居客户。在搜索最近邻居过程中，度量客户之间相似性的方法主要有三种：余弦相似性、相关相似性以及修正的余弦相似性，根据文献[3]的结论，Pearson 相关相似性通用性更好。

定义 2[3]：相关相似性，又称 Pearson 相关相似性。设经客户 i 的评分商品集合为 I_i，客户 j 评分的商品集合为 I_j，共同评分的商品集合用 $Iij=Ii\cap I_j$ 表示，则客户 i 和 j 之间的相似性 sim(i,j)是通过 Pearson 相关系数，即式(1)来度量。

$$sim(i,j)=\frac{\sum_{t\in I_{ij}}(r_{it}-\overline{r}_i)(r_{jt}-\overline{r}_j)}{\sqrt{\sum_{t\in I_{ij}}(r_{it}-\overline{r}_i)^2}\sqrt{\sum_{t\in I_{ij}}(r_{jt}-\overline{r}_j)^2}} \tag{1}$$

其中 $\overline{r}_i=\sum_{t\in I_{ij}} r_i/\left|I_{ij}\right|, \overline{r}_j=\sum_{t\in I_{ij}} r_j/\left|I_{ij}\right|$ 表示客户 i 和客户 j 在共同评分商品集中的平均值。

定义 3：已知数据源 D=(U，I，R)，给定目标客户 g，及其对 I 中商品评分向量 A(g，n)，对于 $\forall$ i∈U，将 sim(g,i)最大的 S 个基本客户 i 组成集合 NSg，则称该集合中的元素为目标客户 g 的最近邻居。

在得到目标客户 g 的最近邻居集合后，若客户 g 已经评分的商品集合为 I_g，尚未评分的集合为 $I_g^c=I\setminus I_g$，当预测客户 g 对商品 k∈I_g^c 的评分时，可以按照定义 4 的方式进行计算。

定义 4：已知数据源 D=(U，I，R)，给定目标客户 g，及最近邻居集合 NSg，则客户 g 对商品 k 的预测评分 $\hat{r}_{gk}$ 为：

$$\hat{r}_{gk}=\overline{r}_g+\frac{\sum_{i\in NS_g} sim(g,i)\times(r_{ik}-\overline{r}_i)}{\sum_{i\in NS_g}\left|sim(g,i)\right|} \tag{2}$$

其中，$\overline{r}_g$ 和 $\overline{r}_i$ 分别表示客户 g 和客户 i 对商品评分的平均值。

2 考虑时间和价格的协调过滤方法

2.1 权重因子的设定

不同客户的兴趣偏好随着时间动态漂移，每个具体客户可能不是在同一时间段内对相同商品产生兴趣，因此利用不同客户在不同时间段内的评分值来寻找最近邻居显然是不合理的。另外 web 客户的购买行为具有时尚型，太长时间以前的兴趣爱好对最近的购买行为缺乏参考价值，因而应该在同一时间段或相近时间段内比较目标客户与各个基本客户对商品评分之间的相似性。为更具一般性，认为距离目前的时间越遥远，推荐价值越小。故而引入时间权重因子 wt。

$$w_t(i,j)=\alpha^{-t_{ij}}\quad(\alpha>1) \tag{3}$$

其中 t_{ij} 表示客户 i 对商品 j 评分的时间，该时间以当前时间为 0 起点，距离现在越远，t_{ij} 取值越大。是 t_{ij} 的单调减函数，且满足 $\alpha^0=1$，$\alpha^{-\infty}=0$，因为 $t_{ij}\in[0,\infty)$，进而∈(0,1],正好满足了要求。

特别是当 α→1 时，$w_t(i,j)\equiv 1$，此时相当于没有考虑时间的影响。并且随着的增大，时间因素的衰减越明显，当 α→∞时，只有当前时间的评分起作用，所有过去的评分都予以忽略，这一规律可以从图 1 看出。

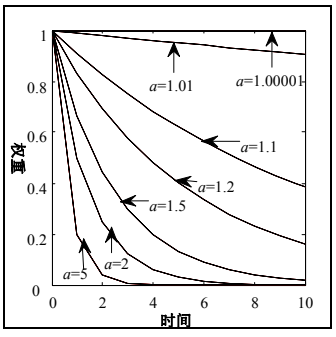

图 1　时间权重因子

衡量商品价值的最重要尺度是价格，客户愿意购买高价格的商品处于兴趣的原因要比购买低价格商品处于兴趣的原因更鲜明，自然高价产品交易后的评分比低价产品的评分更能表明客户的兴趣。因而需要引入价格权重因子 wp。

$$w_p(i,j)=\frac{\beta-\beta^{1-p_{ij}}}{\beta-1}\quad(\beta\neq 1) \tag{4}$$

其中 p_{ij} 表示客户 i 购买产品 j 的价格，令 $p_{ij}\in[0,1]$，在实验前将所用产品的价格除以商品集中的最高价格，进而标准化为[0,1]。该权重因子是 pij 的单调增函数，且满足 $w_0(\cdot)=0$；$w_1(\cdot)=1$，即 $w_p\in[0,1]$。

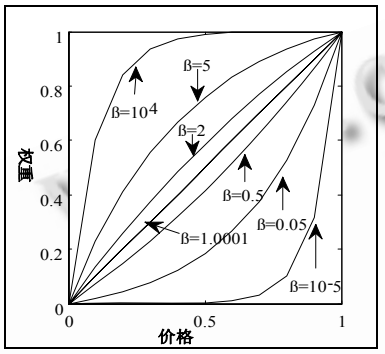

图 2 价格权重因子

当 $\beta\to 1$ 时，$w_p(i,j)=P_{ij}$，即以标准化后产品的价格为权重，当 $\beta\to\infty$ 时，对于 $w_p\neq 0$，$w_p(i,j)\equiv 1$，即相当于没有考虑价格的影响。这一规律可以从图 2 看出。

综合考虑时间和价格因素，得到式(5)所示的综合权重因子 w(i,j)。

$$w(i,j)=\lambda w_t(i,j)+(1-\lambda)w_p(i,j)\quad(0\leqslant\lambda\leqslant 1) \tag{5}$$

其中参数 λ 表示在该商品推荐中时间因素和价格因素的相对重要程度，当 $\lambda=\frac{1}{2}$ 时表示，两者同样重要，当 $0\leq\lambda<\frac{1}{2}$ 表示价格因素更为重要，当 $\frac{1}{2}<\lambda\leq 1$ 表示时间因素更为重要。

2.2 加权协同过滤算法

如何将权重因子加入到传统的协同推荐算法中，文章提供两种方法，分别为考虑时间和价格的最近邻居选择推荐算法和考虑时间和价格的评分预测推荐算法。

2.2.1 方法一：考虑时间和价格的最近邻居选择法

借鉴文献[5]的思路，对寻找最近邻居的相关相似性系数进行改进，对不同客户在不同时间段对不同价格商品的评分值进行加权，即在 Pearson 相关系数公式中用 $w(i,t)\times r_{it}$ 代替 r_{it} 确定客户之间的相似性，如式(6)所示。

其中 $w(i,t)$ 为综合权重因子，r_{it} 为第 i 个客户对 I 中第 t 个商品的评分，其余符号的意义与式(1)相同。

$$sim(i,j)=\frac{\sum_{t\in I_{ij}}[w(i,t)r_{it}-\overline{r}_i][w(j,t)r_{jt}-\overline{r}_j]}{\sqrt{\sum_{t\in I_{ij}}[w(i,t)r_{it}-\overline{r}_i]^2}\sqrt{\sum_{t\in I_{ij}}[w(j,t)r_{jt}-\overline{r}_j]^2}} \tag{6}$$

然后根据预先设定的邻居数 Nn，选择 sim(g,j)最大的前 Nn 个客户作为目标客户 g 的最近邻居。并采用式(2)计算目标客户 g 对任意项 $k\in I_g^c$ 的评分进行预测，然后选择 I_g^c 中排在前 Nc 个的商品作为推荐项。

该方法对应的算法描述为：

1)输入

①客户-商品矩阵，m 个客户对 n 个商品的评分，每个评分对应的时间，每个商品的价格，α、β、λ 参数；

②最近邻居客户个数 Nn，目标客户 g，该客户已评价的商品 I_g，商品评分阈值 ξ。

2)输出

目标客户的感兴趣的商品。

算法过程：

对于任意客户 g 和商品 k，预测客户 g 对第 k 个商品的评价。

①根据式(6)计算目标客户 g 与 m 个客户得相似性 sim(g,j)系数，选取该系数中前 Nn 个对应的客户作为目标客户的邻居客户。

②根据 Nn 个邻居客户在 $I_g^c=I\setminus I_g$ 上的评价分，由式(2)预测目标客户 g 对商品 k 的评价值 $\hat{r}_{gk}$。

③对于预测评分值 $\hat{r}_{gk}>\xi$ 所对应的商品为客户 g 感兴趣的商品。

2.2.2 方法二：考虑时间和价格评分预测算法

传统算法中的评分预测公式(2)本质上就是一种加权平均，其权重就是邻居客户与目标客户的相似程度，如果考虑时间和价格因素后，其修正的公式如式(7)所示。

$$\hat{r}_{gk}=\overline{r}_g+\frac{\sum_{i\in NS_g}sim(g,i)\times w(i,k)\times(r_{ik}-\overline{r}_i)}{\sum_{i\in NS_g}|sim(g,i)|\times w(i,k)} \tag{7}$$

文献[6]和文献[7]在考虑时间权重因素时，也是对预测公式(2)的修正，本文修正公式与文献[6]类似，与

文献[7]有较大区别，这两类修正方法，笔者通过实验证实文献[6]的更有效（鉴于篇幅、且不是本文说明的主要问题，实验从略）。

该方法对应的算法输入与输出与方法一相同，算法过程为：

对于任意客户 g 和商品 k，预测客户 g 对第 k 个商品的评价分。

1) 根据式(1)计算目标客户 g 与 m 个客户得相似性 sim(g,j)系数，选取该系数中前 Nn 个对应的客户作为目标客户的邻居客户。

2) 根据 Nn 个邻居客户在 $I_g^c = I \setminus I_g$ 上的评价分，由式(7)预测目标客户 g 对商品 k 的评价值 $\hat{r}_{gk}$。

3) 对于预测评分值 $\hat{r}_{gk} > \xi$ 所对应的商品为客户 g 感兴趣的商品。

3 实验结果及其分析

本文以某在线购物站点为对象，采用该网站网络数据库中的客户购买评价数据检验本文给出的推荐算法。为了说明算法的偏差大小，引入平均误差 MAE 指标，该指标由 Shardandand&Mases 和 Sarwar 提出，用于度量预测值与实际值之间的偏差。MAE 的定义如式(8)所示。

$$MAE = \frac{\sum_{i=1}^{N} |\hat{r}_i - r_i|}{N} \tag{8}$$

其中 $\hat{r}_i$ 是预测值，r_i 是实际评分，预测商品个数为 N。

在本实验中，共选取有 200 个客户对 15 件商品的评价数据，以其中 180 个客户为基本客户，另外 20 个客户为目标客户，该 20 个客户中，以前 10 件商品评价为已知数据，来预测后 5 件商品的评价值。通过设定参数 α =1.5、β =5、λ =0.5，来验证本文给出的方法一、方法二和传统的协同过滤算法[3]的预测精度。

将目标客户最近邻居个数从 20 增加到 30(间隔为 2)，查看不同的邻居数量大小对算法精确度的影响，结果如图 3 所示。

由实验结果可以看出：整体来说，考虑时间和价格因素的协同推荐算法要比传统的协同推荐算法[3]精确度高，原因是它更能反映客户近期的、投入最大的兴趣爱好；同时也可看到方法一要好于方法二，这是因为方法一在寻找最近邻居时考虑了时间和价格因素，因而得到的是最准确的邻居。虽然在预测公式中没有加入时间和价格的权重因子，但是此时的相关相似系数 sim 已经含有了这部分信息，而方法二在寻找邻居时没有考虑时间因素，得到的邻居便不准确，即便在预测时考虑了这两个因素，改进效果也大大下降。

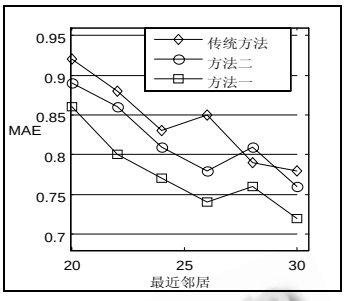

图 3　推荐算法的平均绝对误差 MAE 比较

4 结语

本文首先详细介绍了传统的基于客户协同过滤算法的实现过程，针对传统方法没有考虑商品购买时间和价格因素的缺点，提出了两种改进的方法，这两种方法分别将时间和价格的综合权重因子加入到相似性系数和预测评分公式中进而得到新的计算公式。通过实验表明，整体来说，考虑时间和价格因素的协同推荐算法要比传统协同推荐算法的精确度高，实验另外表明将时间和价格因素加入相关相似性计算公式中更加有效。

参考文献

1 Resnick P, Iakovou N, Sushak M, et al. GroupLens:An open architecture for collaborative filtering of netnews. Proc. of 1994 Computer Supported Cooperative Work Conf. Chapel Hill, 1994. 175-186.

2 Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of use. Proc. of Conf Human Factors in Computing Systems. Denver, 1995. 194-201.

3 Breese JS, Heckerman D, Kadie C, Empirical analysis of predictive algorithms for collaborative filtering. Madison. 14th Conference on Uncertainty in Artificial Intelligence. Wisconsin, Morgan Kaufmann, 1998. 43-52.

4 Dellarocas C. Immuning online reputation reporting systems against unfair ratings and discriminatory behavior. Proc. of the 2nd ACM Conference on Electronic Commerce. Minneapolis, USA, 2000. 150-157.

5 刘枚莲,丛晓琪,杨怀珍.改进邻居集合的个性化推荐算法.计算机工程,2009,35(11):196-198.

6 王岚,翟正军.基于时间加权的协同过滤算法.计算机应用,2007,27(9):2302-2326.

7 丛晓琪,杨怀珍,刘枚莲.基于时间加权的协同过滤算法研究.计算机应用与软件,2009,26(8):120-140.