E-Commerce Letters 电子商务评论, 2016, 5(4), 61-67
Published Online November 2016 in Hans. http://www.hanspub.org/journal/ecl
http://dx.doi.org/10.12677/ecl.2016.54009

Hans汉斯

# Study on Collaborative Filtering Recommendation of Tea with Fusion Category and Score

**Guodong Wu[1,2], Jingxia Li[2], Lijing Tu[2], Fugen Song[1]**

[1]Glorious Sun School of Business and Management, Donghua University, Shanghai
[2]School of Information and Computer, Anhui Agricultural University, Hefei Anhui
Email: 8978850@qq.com

Received: Dec. 1st, 2016; accepted: Dec. 17th, 2016; published: Dec. 27th, 2016



## Abstract

**As one of the main content of the current research and application in the electronic commerce field, personalized recommendation will be the inevitable development in the future of tea e-commerce applications. The existing personalized recommendation technology regarded tea as common goods recommending to consumers, but ignored the agricultural characteristics. And aimed at the problems such as the few recommendations led by the traditional personalized recommendation only considering consumer ratings, this paper proposed a method of the construction of tea consumer network fusing tea category similarity. The experiment on the basis of real data sets indicated that, the new recommendation method improves the personal recommendation accuracy for the tea products, having a certain significance to realize personal marketing of tea e-commerce.**

## Keywords

**Tea, Category, Score, Collaborative Filtering, Recommendation**

# 融合类别与评分的茶产品协同过滤推荐研究

**吴国栋[1,2]，李景霞[2]，涂立静[2]，宋福根[1]**

[1]东华大学旭日工商管理学院，上海

文章引用: 吴国栋, 李景霞, 涂立静, 宋福根. 融合类别与评分的茶产品协同过滤推荐研究[J]. 电子商务评论, 2016, 5(4): 61-67. http://dx.doi.org/10.12677/ecl.2016.54009

[2]安徽农业大学信息与计算机学院，安徽 合肥
Email: 8978850@qq.com

收稿日期：2016年12月1日；录用日期：2016年12月17日；发布日期：2016年12月27日

## 摘 要

**个性化推荐作为当前电子商务领域研究的热点之一，也是未来茶产品电子商务发展的必然趋势。论文针对现有个性化推荐中将茶叶作为普通商品推荐，没有考虑其自身农产品特性，同时对传统协同过滤推荐技术仅考虑消费者评分导致的推荐精度不高等问题，提出将评分与商品类别相融合的方式实现对茶叶的协同过滤推荐。通过真实数据集上的实验表明，新的推荐方法提高了茶产品个性化推荐的准确度，对实现茶叶电子商务的个性化营销具有一定的意义。**

## 关键词

**茶叶，类别，评分，协同过滤，推荐**

## 1. 引言

近年来，茶叶电子商务发展迅速，根据相关资料显示，电子商务作为一种新兴的茶叶销售方式和渠道模式，发展极为快速。如 2015 年仅阿里零售平台茶叶销售额为 88 亿元[1]，预计到 2016 年底，全国茶叶电商的销售总额将超过 160 亿元，越来越多的茶叶企业涉足电子商务。据国家茶叶产业技术体系产业经济研究室调研结果显示，目前 64.4%的茶叶企业开展了茶叶电子商务，20.5%茶企计划开展[2]。特别是随着移动互联网的普及，茶叶移动电商交易占比稳步提升。该调研数据显示，已开展了电子商务的茶企中有 91.5%的企业涉足移动端业务，而计划开展电子商务的茶企中有 93.3%的企业倾向于选择移动端。伴随着茶叶电子商务规模的增大，网上的茶叶商品信息越来越多，信息的泛滥反而使用户无从选择，这就是信息超载(information overload)问题。因此，在茶叶电子商务中引入个性化推荐技术，是未来发展的必然趋势。当前的电商个性化推荐中，只是将茶叶作为普通商品来推荐，忽视了茶叶本身的农产品特性，导致推荐的效果并不理想，制约了茶叶电子商务的发展。如何结合茶叶商品的区域性、时间季节性等特性，更好发现茶叶消费者偏好，提高茶叶电子商务的个性化推荐水平，这将是适应当前大数据时代茶叶精准营销的需要[2]。本文在对传统协同过滤推荐技术进行深入研究的基础上，针对其仅考虑消费者评分进行推荐所导致推荐精度不高问题，结合茶叶本身地域性特征，提出了融合茶叶类别与消费者评分，实现茶叶电子商务个性化协同过滤推荐方法。通过真实数据集上的相关实验，表明了本文所提出方法的有效性。

## 2. 相关研究

推荐系统按照所使用的数据来分类，可以分为协同过滤推荐、内容推荐和社会化推荐等[3]。其中，协同过滤推荐是当前推荐系统中应用最为广泛、最有效的方法之一，在亚马逊、阿里巴巴、京东等主流的电商平台中都得到充分的运用。协同过滤推荐系统的实质就是一个评分预测过程[4]，其关键问题就是用户或项目相似度的计算。目前，主流协同过滤推荐分为两类：基于用户的协同过滤推荐和基于项目的协同过滤推荐。基于用户的协同过滤推荐根据用户对项目的评分矩阵，计算用户之间的相似度，找出目标用户的最近邻居集合后，对最近邻居集合进行加权，从而产生目标用户的推荐集。此类算法能够有效

地使用其他相似用户的反馈信息，为用户产生推荐。但是由于用户涉及的信息量相当有限，用户对项目的评分相对稀少，造成评分矩阵相对稀疏，数据冷启动问题严重，难以找到准确的相似用户集，导致仅使用少量评价数据不可能产生精确推荐，大大降低了推荐系统有效性。基于项目的协同过滤推荐根据对用户已评分项目中相似项目的评分进行预测，从某种程度上减少了评分矩阵稀疏性和冷启动问题对推荐质量的影响。虽然项目间相似性相对稳定，但用户的喜好和兴趣是不断变化的，推荐集覆盖率较低，此类算法也没有提出有效解决这一问题的方法，用户对推荐的满意度较低。

## 3. 融合茶叶类别与评分的综合相似度计算

### 3.1. 相似性计算方法

传统的相似度有皮尔逊相关系数法、向量余弦法、调整的向量余弦法、约束的皮尔逊相关系数法、斯皮尔曼相关系数法等[5]，在不同的应用领域中，选取不同的相似度计算方法。通常利用用户—项目评分矩阵计算用户之间的相似性，采用较多的是皮尔逊相似性计算方法，其也是一种经典的相似性计算方法。设茶叶电子商务中茶叶商品有$s$个，用户个数为$n$个，建立用户－项目评分矩阵$\boldsymbol{M}$：

$$\boldsymbol{M}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1s} \\ r_{21} & r_{22} & \cdots & r_{2s} \\ \cdots & \cdots & & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{ns} \end{bmatrix} \tag{1}$$

其中，$r_{ij}$ 表示用户 $i$ 对项目 $j$ 的评分，如果用户没有购买过某一种产品，或者购买后没有进行评级，则相应的 $r_{ij}$ 为空，在传统的皮尔逊相似性计算方法中，通常假设未评级的项目评分为 0。根据上述矩阵 $\boldsymbol{M}$，利用皮尔逊相关性算法可以完成用户的相似度计算[4]：

$$\operatorname{sim}(x,y)=\frac{\sum_{s\in S_{xy}}\left(r_{x,s}-\bar{r}_x\right)\left(r_{y,s}-\bar{r}_y\right)}{\sqrt{\sum_{s\in S_{xy}}\left(r_{x,s}-\bar{r}_x\right)^2\sum_{s\in S_{xy}}\left(r_{y,s}-\bar{r}_y\right)^2}} \tag{2}$$

其中，$S_{xy}$ 指的是用户 $x,y$ 共同评级的项目集合，$r_{x,s}$ 指的是用户 $x$ 对项目 $s$ 的评分，$\bar{r}_x$ 表示的是用户 $x$ 的平均评分。

### 3.2. 茶叶消费者评分相似性

传统的皮尔逊相似性计算方法将矩阵中的缺省值假设为 0，通常情况下，茶叶电子商务中用户—项目矩阵中存在大量的缺省值。简单的将这些缺省值固定为 0，可能与实际情况出入很大，因此这种方法不能从根本上解决用户评分数据稀疏性问题。在相似性计算中，有用户相似性，同时也有项目的相似性，通过计算未评分项目与用户已经评分项目的项目相似性，然后预测未评分项目的分数，利用计算得到的分数填补矩阵中的缺省值，可以有效解决矩阵稀疏性问题，同时保证了后续计算的准确性[6]。

设计算用户 $i$ 和 $j$ 的相似性，利用皮尔逊公式，应该得到 $i$ 和 $j$ 评分项目的并集 $U_{ij}$，用户 $i$ 的项目评分集合为 $I_i$，用户 $j$ 的项目评分集合为 $I_j$，则：$U_{ij}=I_i\bigcup I_j$。然后分别预测两个用户在该并集中的未评分项目，以用户 $i$ 为例，设用户 $i$ 未评分项目集合为 $N_i$，对任意一个项目 $a\in N_i$，预测用户 $i$ 对项目 $a$ 的评分 $P_{i,a}$。

首先计算项目 $a$ 与其它项目之间的相似性，这也是本算法最为关键的一步，同用户之间的相似性计算类似，采用皮尔逊相似性计算方法计算项目之间的相似性。然后，找到相似性最高的若干项目作为项目 $a$ 的邻居集合，考虑到计算的方便，邻居集合取 6 个，邻居集合 $M_a=\{I_1,I_2,I_3,I_4,I_5,I_6\}$。第三步，利

用项目评分预测公式计算用户 $i$ 对项目 $a$ 的评分 $P_{i,a}$。

$$P_{i,a}=\frac{\sum_{n\in M_P}\mathrm{sim}_{a,n}\times R_{i,n}}{\sum_{n\in M_p}\left(\left|\mathrm{sim}_{a,n}\right|\right)} \tag{3}$$

通过计算，可以使用户 $i$ 和 $j$ 在项目集合 $U_{ij}$ 上均有评分，弥补了矩阵稀疏性带来的计算不准确问题。对任意项目 $w\in U_{ij}$，用户 $i$ 对项目 $w$ 的评分：

$$R_{i,w}=\begin{cases}r_{i,w}, & \text{if } w \text{ rated } U_i\\ P_{i,w}, & \text{if } w \text{ not rated } U_i\end{cases} \tag{4}$$

最后利用皮尔逊公式计算用户 $i$ 和 $j$ 的相似性。

### 3.3. 茶叶类别相似性

利用项目相似性预测用户未评分项目的评分，从而填补矩阵中的缺省值。这种方法的核心是项目相似性的计算，传统的相似性计算完全的依赖评分数据，而忽略的项目自身的分类属性[7]。在实际的茶叶电子商务中，通常利用茶叶的特征属性对茶叶进行细致的分类，本文结合传统分类和电子商务中茶叶产品分类方法，将茶叶分为六大类[8]。根据这样的划分，可以把所有的茶叶商品形成一个倒立的树，这个树一共具有六层。以绿茶分类为例，部分地绘制茶叶分类图，如图 1 所示。

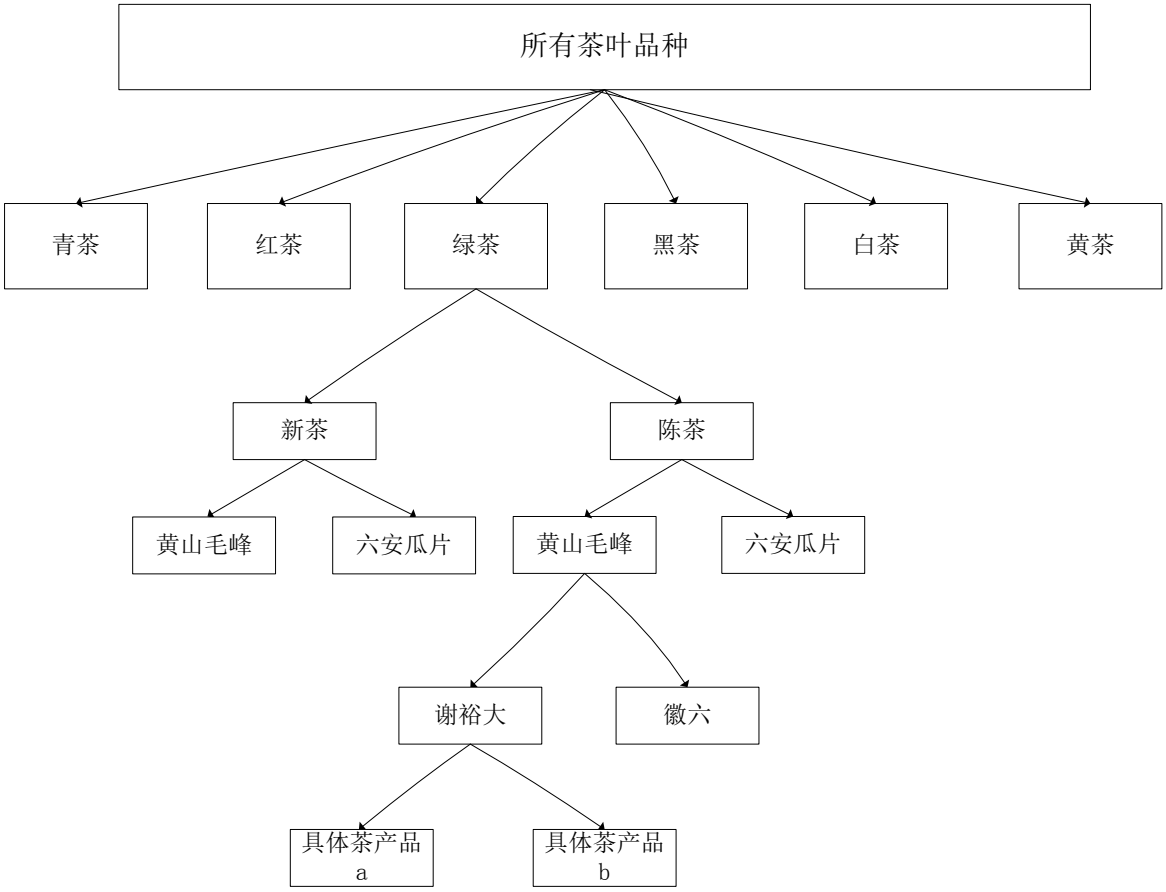


**Figure 1.** Classification of some tea
**图 1.** 部分茶叶分类图

将茶叶的类别属性融入到后续的相似度计算中，可以提高评分预测的准确性。结合电子商务中茶叶的具体分类，可以采用以下做法计算项目的类别相似性：首先根据茶叶项目类别树，定义树总层数为 $H$，在茶叶项目类别树中，一共有 6 层；然后找到项目 $i, j$ 共同所属的类，即公共类，项目 $i, j$ 可能会有多个公共类，称距离项目根节点最远的类为 $i, j$ 的最近公共类，这样两个项目的最近公共类只有一个，定义最近公共类为 $H(i, j)$。在茶叶分类中，每一层的分类依据分别是：色泽分类，生鲜性分类，产地分类，品牌分类。可以采用一种较为简单的判断项目类别相似性的方式：

$$\mathrm{sim}(i, j) = \begin{cases} 0 & (H(i, j) = 1) \\ \dfrac{H(i, j)}{H} & (H(i, j) > 1) \end{cases} \tag{5}$$

利用上述公式，可知，当最近公共类分别为色泽分类，生鲜性分类，产地分类，品牌分类的是，项目类别相似性分别为：$\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$。

### 3.4. 茶叶商品综合相似性

综上，可以得出项目的综合相似性应该结合项目的评分相似性与项目的综合相似性，是一个加权的组合值，定义项目的综合相似性为 $\mathrm{sim}_c(i, j)$，项目的类别相似性为 $\mathrm{sim}_g(i, j)$，项目的评分相似性为 $\mathrm{sim}_m(i, j)$。即得到最后的项目综合相似性：

$$\mathrm{sim}_c(i, j) = (1 - \alpha)\mathrm{sim}_g(i, j) + \alpha \mathrm{sim}_m(i, j) \tag{6}$$

其中，$\alpha$ 为权重系数，文献[9]指出在计算项目综合相似性过程中，非评分相似性对最终结果影响要弱于评分相似性，通常情况下，非评分相似性取 0~0.5 的中间值 0.3 比较合适。在本文中，取项目类别相似性为 0.3。

## 4. 基于综合相似性的茶产品协同过滤推荐算法

① 利用公式(6)分别计算目标用户 $U$ 与其他用户中未共同评分茶产品的相似性；
② 利用公式(3)分别计算目标用户 $U$ 与其他用户中未共同评分茶产品的预测评分；
③ 将新的预测评分补充填入用户－项目评分矩阵中(1)对应的缺省部分；
④ 利用公式(2)计算目标用户与其他用户的相似性；
⑤ 根据最近邻方法，选出 $K$ 个和目标用户 $U$ 最相似的用户；
⑥ 将目标用户 $U$ 没有购买而 $K$ 个用户已购买过的茶产品，推荐给 $U$。

## 5. 仿真实验与分析

为了验证本文所提出算法的性能，在 Intel(R)Core(TM)i7-3770 CPU@3.40GHZ 处理器、4GB 内存的台式机上进行相关实验。实验的数据集选取的是某电子商务网站 2013 年 4 月份到 8 月份期间部分茶叶商品交易数据。去除不活跃用户后，实际数据集由 350 个用户、485 种茶叶商品、5000 条评分记录构成，用户的评分介于 1~5 之间的整数，评分数据越高表明用户对该种茶叶产品喜欢程度越高。算法中的最近邻用户数 K 取值为 10，实验主要采用对茶叶商品的推荐的平均正确率和平均绝对误差(Mean Absolute Error, MAE)来度量推荐的质量。平均正确率表示预测的正确结果与实际所推荐的商品数量所占的比例，平均绝对误差 MAE 是通过比较预测值与用户实际的评分值之间的偏差来衡量预测结果的准确性，MAE 越小，表明推荐质量越高。平均正确率定义为：

$$\text{平均正确率} = \text{正确推荐茶叶商品数量} / \text{所有推荐茶叶商品数量} \tag{7}$$

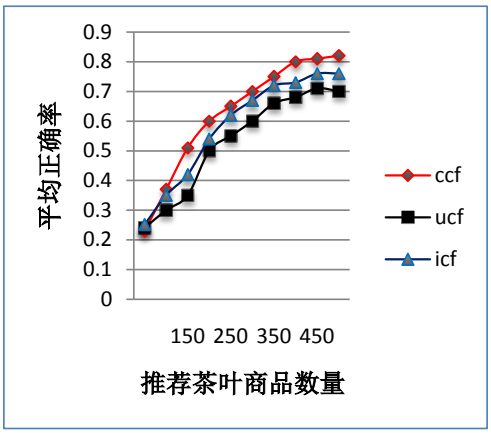


**Figure 2.** Average correct rate of three different recommendation algorithms
**图 2.** 三种不同推荐算法平均正确率

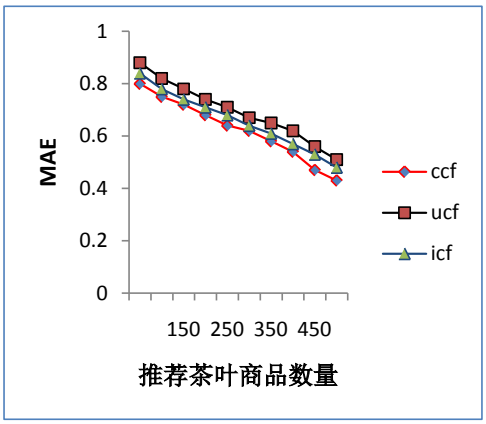


**Figure 3.** Average absolute error of three different recommendation algorithms
**图 3.** 三种不同推荐算法的平均绝对误差

设预测的用户评分集合表示为$\{p_1, p_2, \cdots, p_n\}$，对应的实际用户评分集合为$\{q_1, q_2, \cdots, q_n\}$，则平均绝对偏差 MAE 被定义为：

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |p_i - q_i|}{n} \tag{8}$$

为了验证本文提出的茶产品综合相似性协同过滤推荐算法(CCF)的推荐效果,将其与直接基于用户的协同过滤推荐算法(UCF)和基于项目的协同过滤推荐算法(ICF)三种方法进行对比，结果如图 2 所示。

由图 3 可以看出，在三种不同的推荐方法中，本文所提出推荐方法 CCF 的平均正确率要好于其他两种方法 UCF 和 ICF，并且随着所推荐茶叶商品数量的增加，优势也越来越明显。

## 6. 结束语

随着农产品电子商务的迅猛发展，个性化推荐技术必将在该领域得到更为广泛的应用。茶叶作为一种特色农产品，在网络营销方面有着自身的优势。如何让消费者购买到自己满意的茶产品，如何让商家向消费者推荐其满意的茶产品，个性化推荐有着重要的意义。本文在对传统协同过滤推荐技术进行分析

的基础上，结合茶产品自身的特性，提出了一种融合评分与茶产品类别相似性的协同过滤推荐方法，通过实验验证了该算法与传统协同过滤算法相比，对茶产品推荐有着更高的准确度，论文研究成果将对推动茶产品及其它农产品电子商务的发展有着一定的指导意义。

## 基金项目

国家自然科学基金(31671589)；安徽省科技攻关重点项目(1501031082)。

## 参考文献 (References)

[1] 阿里研究院. 2015 茶叶电商微报告[Z/OL]. http://www.aliresearch.com/Blog/Article/detail/id/20884.html, 2016-3-30.

[2] 中国茶叶学会. 2015 年度中国茶叶产销形势报告发布[Z/OL]. http://www.puercn.com/chayenews/cyzx/100626.html, 2016-1-10.

[3] 郭磊, 马军, 陈竹敏, 姜浩然. 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): 219-228.

[4] 陆春, 洪安邦, 宫剑. 基于 PSO 的协同过滤推荐算法研究[J]. 计算机工程与应用, 2014, 50(5): 101-107.

[5] 荣辉桂, 火生旭, 胡春华, 莫进侠. 基于用户相似度的协同过滤推荐算法[J]. 通信学报, 2014, 35(2): 16-25.

[6] 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): 25-29.

[7] 黄霞. 基于用户属性和项目类别协同过滤算法[J]. 计算机与数字工程, 2012, 40(10): 5-7.

[8] 古能平. 关于茶叶分类的几点认识[J]. 消费导刊, 2008(17): 204+229.

[9] 范波, 程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学, 2012, 39(1): 23-26.