基于商品基因和遗传算法的个性化推荐系统①

张 浩

(淮阴工学院 交通工程学院，淮安 223003)

摘 要：为解决“新用户”和“稀疏性”问题，引入商品基因的概念，通过将商品基因库、用户历史行为库、用户在线浏览内容及邻近用户行为数据耦合，形成用户偏好度候选集的兴趣模式抽取模块，然后利用改进的遗传算法优化模块进行模式选取与聚合，完成最优邻居的选择，最后经由推荐模块产生最终的推荐项目集。实验结果表明，提出的算法提高了推荐的准确度和覆盖面。

关键词：商品基因；遗传算法；混合；推荐算法；推荐系统

Personalized Recommendation System Based on Commodity Gene and GA

ZHANG Hao

(Department of Transportation Engineering, Huaiyin Institute of Technology, Huai’an 223003, China)

Abstract: To solve the problems of "new user" and "sparseness", we introduce the concept of commodity gene. Through coupling the commodity gene database, users’ purchasing historical records, users’ content of online browsing and the data of neighbors’ behavior, we can form the module of candidated sets of customer preferences, and then use genetic algorithm which has be improved to make the selection and polymerization to the model, so that we can complete the best selection of neighbors. Finally we can get the recommended sets according to the recommended module. Experimental results show that the algorithm we suggested can improve the accuracy of the recommendation and achieve good quality of recommendation.

Key words: commodity gene; genetic algorithm; Mixed; recommendation algorithm; recommendation system

1 引言

推荐算法是个性化推荐系统的核心，关系到推荐质量的好坏，一般主要包括协同过滤、关联规则和基于内容的推荐算法等[1]。协同过滤算法基于目标用户的邻居用户的评价资料，其存在着冷启动、特殊用户和稀疏性问题，在推荐效率和推荐质量上存在一些缺陷[2]。关联规则推荐算法根据所有用户的购买习惯建立的产品之间的关联，其中没有考虑到目标用户特有的个性[1]，推荐缺乏针对性，随着站点结构、内容的复杂度和用户人数的不断增加，现有的推荐算法难以表现出良好的性能。针对这两种推荐算法存在的问题，本文引入商品基因的概念，通过将商品基因库、用户历史行为库、用户兴趣偏好集及最优邻近用户集结合起来，提出一种基于商品基因和遗传算法的个性化推荐算法。该算法可以弥补现行推荐算法推荐精度不高、推荐效率低、新上市或购买率较低的商品不能及时推荐给客户的不足，同时利用遗传算法对顾客兴趣度偏好候选集进行学习，得到最佳邻居用户，使得推荐商品的覆盖率得到一定的提升，推荐效果明显改善。

2 基于商品基因和遗传算法的混合推荐算法

2.1 推荐问题的形式化描述

推荐问题可以用以下形式化的方法加以阐明[3]：

C 表示所有用户(Customer)的集合，S 表示所有可能被推荐的商品项目(Subject)集合。函数 u 作为度量 s

① 收稿时间:2011-05-27;收到修改稿时间:2011-07-07

对 c 的效用函数，u 的定义为：

$$u: C \times S \to R \quad (1)$$

R 是所有用户的评价集合。对于用户空间中的每一个用户，我们的目的是从商品项目空间中寻找能使用户效益(代表用户的满意度)最大化的商品项目，即：

$$\forall c \in C, S_c = \max u(c, s), s \in S \quad (2)$$

2.2 算法模型

2.2.1 商品基因

商品基因的选取建立在客观、可识别描述的基础上，其与商品基因库中的特征基因相匹配，例如，商品种类、品牌、价格、颜色、对商品种类的点击次数等。用向量来表示为：$P(n)=(f_1, f_2, \cdots, f_m)$，$P\_(no)=(f_{11}, f_{12}, \cdots, f_{ik}, \cdots, f_{im})$，n 为某一实体商品，如笔记本电脑，$f_m$ 为笔记本电脑的第 m 项基因的编号。P_no 为产品编号，f_{ik} 为基因 i 的第 k 项特征基因值。每一个基因值用二进制来表示，例如，若某产品具有某项特征基因值则其值为 1，反之为 0。

2.2.2 用户对商品基因的偏好度

根据用户某段时间内的个人交易纪录并结合商品基因库，分析其对商品基因的偏好程度，根据偏好程度对这些特征基因赋值，得到当前用户对于特征基因的偏好度，作为系统推荐依据。其顾客偏好商品特征基因的产生方法如下：

$$UIP_k^i = \frac{S_i}{\sum_{j=1}^{n} S_j} \quad (3)$$

其中 UIP_k^i 为顾客 k 对商品特征基因 i 的喜好程度，通过对顾客 k 在一段时间内所购买商品的部分主要商品特征基因进行统计，描述 k 对特征基因 i 的偏爱度 UIP_k^i 即为客户购买商品中某项特征占全部购买商品包含的总特征的比值，比值越大表示该特征越受顾客喜爱。n 表示统计商品特征基因的数量，S_i 为某时间段内客户 k 购买商品包含某项特征基因的数量。以表 1 为例，则 $UIP_{001}^1=0.624, UIP_{001}^2=0.624, UIP_{001}^3=0.624$。

表 1 顾客 001 产品购买表

C_no	P_no	f_{11}	f_{12}	f_{23}	…	f_{43}
001	A001	1	0	1	…	0
001	A002	1	1	0	…	0
001	A005	1	0	1	…	1

通过对客户购买日志的分析，由公式 3 计算出其对各个特征的偏好度 UIP_k^i，然后根据偏好度的值进行排序，出于对计算精度的考虑，在实际推荐时可以只取偏好度排在前 i 位的特征用于推荐，如按照偏好度值进行计算排序的结果为:品牌，价格，重量，…，相应特征的偏好度依次 0.418，0.624，0.237, …。求得顾客对某类产品特征的偏好度后，我们可产生顾客的产品喜好模式。

$$UIP_k = (UIP_k^1, UIP_k^2, \cdots, UIP_k^n) \quad (4)$$

最后将顾客的产品特征偏好模式、顾客交易纪录和顾客个人档案经由整合计算并建立顾客偏好度偏好模式候选集。

$$C_k = (UIP_k, CI_k) \quad (5)$$

$$CI_k = (CI_k^1, CI_k^2, \cdots, CI_k^n,) \quad (6)$$

C_k 即为顾客 k 的顾客偏好度模式公式，其中包含 UIP_k 与 CI_k。UIP_k 为用户 k 的产品喜好模式，CI_k 为顾客 k 的基本资料与交易纪录，见表 2。顾客交易纪录见表 3，其 C_no 为顾客编号对应于产品编号 P_no，以及所交易的产品数量 B_qty、产品交易金额 B_mny 和交易日期 B_day。表 4 为顾客个人档案，其为存放顾客个人的基本资料。

表 2 顾客偏好度候选集

	1	2	3	…	15	…
C_no	UIP_{001}^1	UIP_{001}^2	UIP_{001}^3	…	UIP_{001}^{15}	…
001	0.624	0.237	0.418	…	0.132	…

表 3 顾客交易记录表

C_no	P_no	B_qty	B_mny	B_day
001	A001	1	260	2008/11/03
003	A002	3	850	2009/02/06
005	A005	2	570	2008/10/07

表 4 顾客注册信息表

C_no	C_age	C_lv	C_sex
001	30	2	0
003	27	3	1
005	20	1	0

2.2.3 基于 GA 的最优邻居选择

当顾客进行浏览和购买行为时，记录顾客当前正在浏览的产品类别，找出该产品类别中符合顾客偏好

的产品，以进行推荐[4-5]。首先利用遗传算法计算顾客的特征权重，再利用欧几里德距离公式找出与顾客具有相同喜好的邻近群体，同时由特征权重来调整以更符合顾客的偏好度。其中的重点主要在于进化计算和模式匹配的工作。其主要的步骤如下：

1) 样本选取

由于顾客偏好模式库的数据都非常的庞大，如果将其全部加以运算处理则会不经济且没有效率。因此，大部分的系统都是随机选择部分的样本进行运算处理，而种群大小要由问题特性或时间成本来决定，因为当样本数越大，达到收敛的时间就越长。

2) 模式聚合

由于不同顾客的年龄、性别、职业等不同，因此顾客在其对产品喜好的特征权重上就有其不同的重要性。本文利用遗传算法来计算其特征权重。对于顾客 K 其特征权重为 W(K)如下所示：

$$W(K)=(W_{f1},W_{f2},\cdots,W_{fi}) \tag{7}$$

W(K)为顾客 K 的特征权重集合，为产品的特征权重其基因状态以二进制的方式表示，每一个体包含 i 个基因。

pred(A_i,j)为本文所使用的适应性函数。首先我们将数据分成训练组与测试组。对于每一个测试组的顾客(例如顾客 A)，我们预测他对产品 j 的喜好如下：

$$pred'(A,j)=\sum_{k=1}^{n} similarity(A,k)\,vote(k,j) \tag{8}$$

$$pred(A,j)=\begin{cases}1 & if \quad pred'(A,j)\geq 0,\\ -1 & otherwise\end{cases} \tag{9}$$

其中 n 为邻居数，vote(k,j)为顾客 k 对产品 j 的购买喜好，若顾客 k 曾经购买此产品则其值为 1，否则为-1。 Similarity(A, k)为顾客 A 与 k 的相似度，其公式如下所示：

$$Dist(A,k)=\sqrt{\frac{\sum_{i=1}^{n} W_{fi}\times(C_A^i-C_k^i)^2}{\sum_{i=1}^{n} W_{fi}}} \tag{10}$$

$$Similarity(A,k)=1-Dist(A,k) \tag{11}$$

其中 A 表示线上顾客，k 为模式选择处理的顾客且$k\neq A$，W_{fi}表示顾客 A 的偏好度的特征权重，当W_{fi}为 0 时则此特征将被忽略，C_A^i为线上顾客 A 偏好度中的特征 i 的喜好程度，C_k^i为顾客 k 偏好度中的特征 i 的喜好程度。在计算 Dist(A,k)时，C_A^i与C_k^i必须先做正规处理。 依据预测喜好，我们接着计算预测误差如下：

$$diff(A,j)=|pred(A,j)-vote(A,j)| \tag{12}$$

本文随机选取若干个产品，然后计算每个测试组的顾客的平均预测误差 diff(A)。此平均预测误差函数可视为顾客的适应性函数。若某顾客的适应函数值大于事先定义的临界值，则其特征权重将须进一步的演化，然后重新计算其适应函数值直到低于临界值为止。

3) 最佳邻居选择

得到顾客的最佳特征权重后，我们可以由公式 11 来求得某顾客与其他顾客间的相似度，并从中选取最相似的 N 个邻居。

2.2.4 根据偏好度和最优邻居组进行商品推荐

根据前面所得出的用户个体偏好度与商品特征数据库进行匹配，生成符合用户兴趣的推荐组，完成个性化的推荐。同时利用 GA 完成对用户近邻居的发掘，来完成社会化的推荐过程。对于每一个顾客，采用的策略是推荐那些他未曾购买过的产品，但曾经被其最相似的 N 个邻居所购买过。若一个产品被购买的次数越多，则应有较高的机会来推荐。这些产品依其推荐分数来排序，其推荐分数的计算如下：

$$rec(j,K)=\sum_{i\in TN(K)} similarity(K,i)\times pc(i,j) \tag{13}$$

$$pc(i,j)=\begin{cases}1 & 假如用户\,i\,曾购买产品\,j,\\ 0 & 其他\end{cases} \tag{14}$$

rec(j, K)为产品 j 对于顾客 K 的推荐分数，TN(K)为顾客 K 的最相似的 N 个邻居。我们可采用固定数目的推荐方式或设定一个最小推荐分数来筛选。

3 实验结果及分析

本实验采用的是 M 公司提供的数据集，整个实验数据集需要进一步划分为训练集和测试集，为此，我们引入变量 x 表示训练集占整个数据集的百分比。例如，x＝0.6 表示整个数据集的 60%作为训练集，40%作为测试集。在本文的所有实验中，均采用 x=0.6 作为实验基础。

实验在用户相似度算法的精度、推荐命中率二个方面同其他的算法进行比较，验证算法的性能。主要

数据来源有：商品特征数据库，顾客交易记录数据库，用户个人信息数据库，用户偏好模式候选集数据库。随机选择部分样本进行运算处理。使用单点交配其交配率为 0.7，突变率为 0.02，每一基因的编码则使用 8 bits 的 binary 来进行实验。选取 pred’(A,j)为本文所使用的适应性函数，得到顾客的最佳特征权重后，Similarity(A, k)为顾客 A 与 k 的相似度，我们可以由公式 11 来求得某顾客与其他顾客间的相似度，并从中选取最相似的 N 个邻居。最后利用 13、14 确定产品 j 对于顾客 K 的推荐分数，按照分数值从高到低产生前 i 项推荐商品结果集。

兼顾到最优邻居的数量和质量，设定最近邻居阈值大于 0.5 的做为选择最优邻居的标准，在相同的试验样本和环境下，比较本文提出的 ICRS 算法和 Pearson 算法、Vector 算法在获取邻居用户数量和质量上的能力，如图 1 所示。

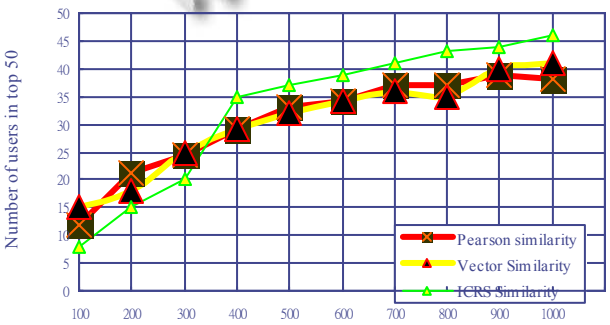


图 1 三种算法下的最优用户邻居情况

为了对推荐结果进行分析，引人推荐命中率这一概念，定义如下：

$$RSP = \frac{\sum_{i=1}^{n} Csum_i}{\sum_{i=1}^{n} Ritems_i} \quad (15)$$

其中，i 为第 i 次推荐，Csumi 代表客户在第 i 次推荐中点击所推荐商品的数量，Ritemsi 表示第 i 次推荐商品的数量，n 为当前客户推荐的次数。

比较传统的协同过滤算法 CF 和本文所提出的 ICRS 算法，根据客户不同数量的购买纪录进行推荐后，统计客户的推荐命中率，所得的结果如图 2 的推荐命中率曲线。

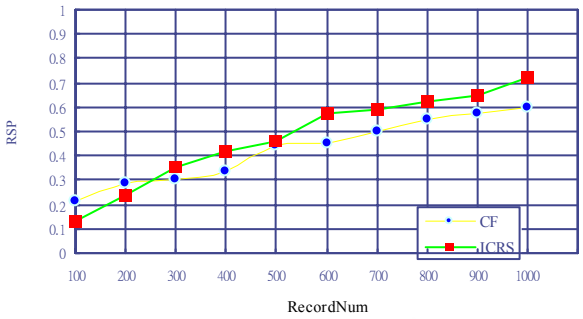


图 2 推荐命中率曲线

由图 2 可见随着分析的购买日志数量和用户数量增加，两种推荐算法的推荐的精度在不断地提高，但是从 200 条之后 ICRS 算法的精度明显要好于 CF 算法。超过 800 条之后，推荐命中率的增加趋势就不十分明显了，这可以作为进行推荐参数设置的一个参考。另外伴随着基础推荐参照组的增多，推荐的精度也有所提高，但是推荐命中率的增幅也比较小，分析其原因，是由于基础推荐参照组的增多，导致了特征范围的扩大，在一定程度上提高了推荐的精度，但同时也扩大了商品类型，在一定程度上又影响了推荐的精度。

4 结语

在整个推荐试验中，按照商品基因进行推荐，同时将客户的历史数据和当前的浏览行为作为推荐依据的一部分，完善了传统推荐的一些不足，比如新商品可以及时推荐给客户，推荐精度有所提高。同时由于不同顾客的年龄、性别、职业等不同，对某一商品基因喜好的特征权重存在差异，这些特征权重受顾客的喜好所影响。本文则利用遗传算法来计算其特征权重，决定个人的最佳特征权重以反应对不同商品基因的喜好程度，有效的降低了上述因素的影响。并且利用遗传算法来决定个人的最佳特征权重以反应对不同喜爱特征的重要性，发掘用户的最优近邻的购物行为模式，对其进行相关推荐，增加了推荐结果的覆盖率，同时准确率也很高。

参考文献

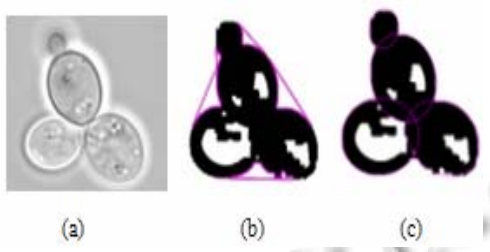
1 余力,刘鲁.电子商务个性化推荐研究综述.计算机集成制造系统,2004,10(10):1306−1313.

2 李雪峰,刘鲁,张曌.基于协同过滤的在线拍卖商品推荐.计

（下转第 166 页）

(a) 原图　(b) 凹点检测图　(c) 椭圆拟合图

图 3-3　实验效果图 2



(a) 原图　(b) 凹点检测图 (c) 椭圆拟合图

图 3-4　实验效果图 3

4　结论

针对酵母菌存在菌体粘连及芽体未与母体分离情况，本文提出了一种基于图像的酵母菌出芽率分析的新算法，利用菌体为椭圆形特征，以边缘和凹点为信息对酵母菌逐一识别。该方法根据粘连程度与芽体分离情况进行判定，能够准确有效的计算出酵母菌出芽率，相对于传统的人工识别方法有明显的改善。

表 1　拟合所得到椭圆的中心和长短轴

椭圆数据	倾斜角	中心	短轴	长轴	面积
4-2-1(上)	72°	(357,91)	77	95	22969
4-2-2(中)	62°	(243,140)	79	103	25550
4-2-3(下)	12°	(171,292)	44	50	6908
4-3-1(上)	356°	(126,125)	57	78	13960
4-3-2(下)	7°	(129,223)	30	36	3391
4-4-1(最小)	34°	(106,319)	32	36	3617
4-4-2(右下)	63°	(235,121)	64	84	16880
4-4-3(中间)	23°	(148,230)	70	82	18023
4-4-4(左下)	94°	(98,123)	70	85	18683

参考文献

1 熊子书.中国酿酒酵母菌的研究.中国食品发酵工业研究所,2002,112(4):23–27.

2 傅蓉.基于凹点搜寻的重叠细胞图像自动分离的算法研究.南方医科大学,2007,43(17):21–28.

3 求是科技.Visual C++数字图像处理典型算法及实现.北京:人民邮电出版社,2006.

4 刘润涛.一种简单多边形凸包的新线性算法.工程图学学报,2002,2:120–125.

5 Yan L, Park CW. New separation alforithm for touching grain kernels based on contour segments and ellipse fitting. Zhejing Univ-Sci, 2011,12(1):54–61.

6 Fitzgibbon A. Pilu M, Fisher RB. Direct least square fitting of ellipses. IEEE Trans. Pattern Anal. Mach. Intell., 1999,21(5): 476–480.

7 Ge P. Research on Segmentation Algorithm of Adhesive Plant Grain Image. 8th Int. Conf. on lectronic Measurement and Instrucents. 2007,2: 927–930.

8 Carter RM. Digital imaging based classification and authentication of granular food products. Meas. Sci. Technol., 2006, 17(2):235–240.

9 彭青玉.木星土星边缘的椭圆拟合.云南天文台台刊,2003,(4):43–46.

（上接第 117 页）

算机工程,2006,(12):18–21.

3 Ko SH. Prediction of Preferences Through Optimizing Users and Reducing Dimension in Collaborative Filtering System. Proceedings of the 17th International Conference on Innovations in Applied Artificial Intelligence. Ottawa, Canada, 2004

4 朱征宇,裴仰军,等.个性化服务中用户近期兴趣视图的生成.计算机工程与设计,2005,(6):237–249.

5 张伟,廖晓峰,吴中福.一种基于遗传算法的聚类新方法.计算机科学,2002,(6):114–116.

6 陶俊,张宁.基于用户兴趣分类的协同过滤推荐算法.计算机系统应用,2011,20(5).

7 肖慧,王立.Web 日志挖掘中的用户识别算法.计算机系统应用,2011,20(5).