第 19 卷第 4 期　　广西科学院学报　　Vol. 19, No. 4
2003 年 11 月　　Journal of Guangxi Academy of Sciences　　November 2003

基于协同过滤技术的 CRM 主动营销模型研究*

CRM Active Marketing Model Based on Collaborative Filtering

覃　华　　苏一丹　　陈　琴
Qin Hua　Su Yidan　Chen Qin

（广西大学计算机与信息工程学院　南宁　530004）
(College of Computer & Information Engineering, Guangxi University, Nanning, 530004)

摘要　叙述在 CRM 中进行主动营销的重要性，提出基于协同过滤技术的 CRM 主动营销模型，对模型中常用的项目相似度计算方法进行详细阐述，然后通过实验证明模型的可行性，并从平均绝对偏差角度说明几种相似度计算方法的性能。

关键词　CRM　主动营销算法　协同过滤技术
中图法分类号　TP274

Abstract　The essentiality of active marketing in the CRM is presented, then an active marketing model based on collaborative filtering is proposed, and the similarity computation methods of this model are stated in detail. Its feasibility is proved by an experiment, and the capabilities of the similarity computation methods are showed in the MAE point.

Key words　CRM, active marketing algorithm, collaborative filtering

高度发达的市场经济，让客户有了更多的选择余地，客户更容易流失。而对电子商务这种新的商务模式而言，如何提高客户忠诚度及保持住客户，如何实现交叉销售等成为电子商务成败的一个关键问题。消费环境的变化，要求企业应进行个性化主动营销。CRM (Customer Relationship Management，客户关系管理) 提供对潜在客户的数据采集和需求验证，对客户的可能喜好动机进行描述。个性化的主动营销推荐系统被用来帮助客户在大量的信息中寻找感兴趣的内容，它体现的“个性化”服务目前越来越为商务网站、电子图书馆等众多领域所接受。因此，在 CRM 中进行主动营销具有潜在的应用价值。

1　协同过滤技术

协同过滤技术也称为面向客户的技术，即协同过滤技术通过分析历史数据，生成与当前客户行为兴趣最相近的客户集，将他们最感兴趣的项目作为当前客户的推荐结果，即 Top-N 推荐[1]，实现个性化的主动营销。

2003-06-27 收稿。
* 广西教育厅教育科学“十五”规划研究课题(2001C108)和广西大学科研基金(X022087)资助。

协同过滤推荐技术是参考客户自身和其他客户的观点来产生对目标客户的商品项目推荐列表。它基于这样一个假设：客户对商品项目的喜好程度用商品项目评分来表达，现已知客户对商品项目 A 的评分，而商品项目 B 的评分未知，如果发现商品项目 A 和商品项目 B 是相似的（相似值反映客户对商品项目的喜好程度），则可参考商品项目 A 的评分来估算客户对商品项目 B 的评分，从而估算出客户对商品项目 B 的喜好程度，由此估算向客户推荐商品项目 B 的成功率。

协同过滤技术是通过使用客户先前的购物爱好及其它相似客户的购物观点，向目标客户推荐新的商品项目或者预测一个商品项目对目标客户的有效性。在一个经典的协同过滤模型中，m 个客户的列表记为：$U=\{u_1,u_2,\cdots,u_m\}$，n 个商品项目的列表记为：$I=\{i_1,i_2,\cdots,i_n\}$，与 U 中的第 i 个客户 u_i 对应的商品项目列表记为 $I_{ui}=\{R_{i1},R_{i2},\cdots,R_{in}\}$，$R_{ik}$ 为 U 中的第 i 个客户对 I 中的第 k 个商品项目的真实评分值。这些评分值可以来自客户的直接评分，也可以来自客户的购物记录或通过日志分析、Web 挖掘等技术获得。客户评分数据可以用一个 $m\times n$ 阶矩阵 $A(m,n)$ 表示，矩阵中第 m 行代表第 m 个客户的相关数据，第 n 列代表第 n 个商品项目，第 i 行第 j 列的元素 R_{ij} 代表客户 i 对商品项目 j 的评分值。注意到：对于任一个客户的评分列表 $I_{ui}\in I$，其中的某个 R_{ik} 可能是一个空值，定义 R_{ik} 的预测估计值为 $P_{i,j}$，它的值落在 R_{ik} 的值域内。存在一个不同的客户 $u_a\in U$（称之为活动客户），则协同过滤算法的任务是发现这 2 个客户实体的相似商品项目列表，由此列表计算出 $P_{i,j}$ 的值或者为客户 u_i 产生一个 N 个商品的推荐列表 $I_r\subset I$，但 $I_r\cap I_{ui}=\Phi$（即为客户产生一个 Top-N 推荐）。

2 基于协同过滤技术的主动营销模型

度量客户 i 和客户 j 之间相似性的方法是：首先得到客户 i 和客户 j 评分过的所有商品项目，然后通过不同的相似性度量方法计算客户 i 和客户 j 之间的相似性，记为 $Sim(i,j)$。协同过滤技术通过访问客户已经投票的商品项目评分数据，计算出商品项目 i 与其它商品项目的相似度，并从中选择相似性最高的 k 个商品项目 $\{T_1,\cdots,T_k\}$ 作最近邻居列表（即最相似的商品项目列表），然后根据最近邻居对待估算商品项目的评分进行预测[2]。

为了找到目标客户的最近邻居，必须度量客户之间的相似性，然后选择相似性最高的若干客户作为目标客户的最近邻居。目标客户的最近邻居查询是否准确，直接关系到整个营销系统的推荐质量，准确查询目标客户的最近邻居是整个协同过滤推荐成功的关键。主动营销模型主要包括两大部分：相似度计算和产生预测估计值。

2.1 项目相似度的计算方法

主动营销算法最关键的一步就是计算商品项目间的相似度，再从中选出最相似的商品项目。计算商品项目 i 和 j 的相似性时，首先找出已经对这 2 个商品项目评分的所有客户，然后根据这些数据计算 i 和 j 的相似度 $S_{i,j}$。相似度的计算方法一般有：余弦相似性法、相关相似性法、修正余弦相似性法[3]。

2.1.1 余弦相似性计算算法

这种算法把 2 个商品项目当作是客户空间上的 2 个 m 维向量，2 个商品项目的相似度由相应 2 个向量的余弦夹角决定。在上述的 $m\times n$ 评分矩阵中，商品项目 i 和 j 的相似度 $Sim(i,j)$ 定义为：

$$S_{i,j} = Sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\| \vec{i} \|_2 \cdot \| \vec{j} \|_2}.$$

分子为 2 个向量的内积，分母为 2 个向量模的乘积。夹角越小，相似度越高。

2.1.2　相关相似度计算算法

在这种算法中，为了计算商品项目 i 和 j 的相似度，首先要找出相关的客户，把这些客户放入数据集 U 中，设 u 是 U 中的一个客户，则此时有：

$$S_{i,j} = Sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - R_i)^2}\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_j)^2}},$$

式中 $R_{u,i}$ 表示 u 号客户对商品项目 i 的评分，$\bar{R}_i$ 是所有客户对商品项目 i 的评分的平均值。

2.1.3　修正余弦相似度计算算法

余弦相似度计算算法有一个不足：计算 i 和 j 的相似度时，没有考虑不同客户在评分尺度上的差异，对此我们可以作出修正，每一个商品项目减去一个客户对所有商品项目评分的平均值，这时有：

$$S_{i,j} = Sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - R_u)^2}\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}},$$

式中 $\bar{R}_u$ 是第 u 个客户对所有商品项目评分的平均值。

2.2　预测计算

主动营销模型最重要的一步就是产生预测。用相似度解决了商品的相似问题后，下一步就要考虑目标客户的评分，并且使用一种技术来获得预测估计值，这里我们介绍 2 种这样的方法[4]。

2.2.1　加权和算法

在这种算法中，要计算客户 u 对商品项目 i 的评分估计值。对于这个客户而言，用与商品项目 i 相似的其它商品项目的评分加权和来计算商品项目 i 的评分值，即：对于一个与 i 相似的商品项目 j，将它的评分值通过相关相似度 $Sim(i,j)$ 进行加权求和。处理过程为：首先通过相似度计算，发现商品项目(如商品项目 1、3、$k-1$、$k+1$、$n-1$) 与商品项目 i 相似，这 5 个商品项目构成了商品项目 i 的相似商品项目集，记为 $D = \{i_1, i_3, i_{k-1}, i_{k+1}, i_{n-1}\}$。不失一般性地对于客户 u，待测的某商品项目 i 的评分值待估算，与商品项目 i 相似的商品项目集记为 $D = \{i_1, i_2, i_3, \cdots, i_N\}$，共有 N 个($N \leqslant$ 总项目数)，客户 u 对这 N 个项目的评分值已存在，记 $R_{u,N}$ 为客户 u 对 D 中第 N 个项目的评分值，$S_{i,N}$ 为 D 中第 N 个项目与待测项目 i 的相似度，则客户 u 对项目 i 的评分预测估计值为：

$$P_{u,i} = \sum_{p=1}^{N}(S_{i,N} \times R_{u,N}).$$

这种方法试图通过相似度模拟客户对项目 i 的评分，这个评分值计算出来后，还要作适当的量化，使它的值落在预定义的值域内。

2.2.2　线性回归法

这个方法类似于权重和法，但它不是直接使用相似项目的评分，而是使用一个基于线性回归模型的近似评分值。实际上，使用余弦和相关性度量计算相似度时，在某种意义上可能会使 2 个向量产生很大的差异(如从欧几里德角度上考虑)，但实际上这 2 个向量可能有高度的

相似性。在这种情况下，使用所谓的相似项目的原始评分值来计算估计值时，可能得到的估计值的可信度差。解决这个问题的基本思想是：计算上式时，不是使用项目 N 的原始评分值 $R_{u,N}$，而是使用它们基于线性回归模型的近似值 $\overline{R}'_{u,N}$，对于目标项目 i 以及它的相似项目 N，通过得线性回归模型为：

$$\overline{R}'_N = \alpha \overline{R}_i + \beta + \varepsilon.$$

这种方法试图通过相似度模拟客户对项目 i 的评分。这个评分值计算出来后，还要作适当的量化，使它的值落在预定义的值域范围内。回归参数 β、ε 由向量决定。本文采用线性回归方法产生预测。

3 实验

以来自 http：//www. eachmovie. com 上的一个网上影片销售系统中的数据进行实验。销售系统提供的实验数据库中有 72916 个客户，1628 部影片，2811983 个数值评分。数据库包括 3 个关键的表：person (ID，age，gender，zip _ code)，movie (ID，name，pr _ url，theater _ status，theater _ release，video _ status，video _ release)，vote (person _ ID，movie _ ID，score，weight，modified)。

从数据库中选取包含有 68000 个有效评分的客户记录集(只选择对 20 部以上影片作出评分的客户)，将选出的数据集一部分作训练数据集，另一部分用作测试数据集。为此，定义一个变量 X，当 $X = 0.8$ 时，表示选出的数据集中，80% 作训练样本，另外的 20% 作测试样本。数据集转换成一个客户项目矩阵 A。从数据库中选取 575 个客户，688 部电影，构成一个 575×688 的矩阵。为评价数据集的质量，引入一个度量 ML 来衡量矩阵 A 的稀疏性，定义为：

$$ML = 1 - \frac{a1}{a2}.$$

式中的 $a1$ 为矩阵中客户作出的评分值的总个数，$a2$ 为矩阵 A 中评分值的总个数，ML 的值越大，矩阵 A 的稀疏性越强。

为了评价算法的性能，引入平均绝对偏差 MAE (Mean Absolute Error) 度量。统计精度度量方法中的 MAE 易于理解，它通过计算测试样本评分的真实值和预测估计值之间的偏差来评价算法的准确性。MAE 越小，算法质量越高。对于测试样本中的一对评分值的真实—预测值对 (p_i, q_i)，取两者的绝对误差作偏差，如 $|p_i - q_i|$。设测试样本中共有 N 对这样的数据对，则定义 MAE 为：

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}.$$

以传统的协同过滤推荐算法作为对照，在传统的协同过滤推荐算法中，分别以余弦相似性和相关相似性作为相似性度量标准，计算其 MAE，然后与本文提出的基于项目评分预测的协同过滤推荐算法作比较，比较结果如图 1 所示。由图 1 可看出回归算法的 MAE 较小。说明使用协同过滤算法预测客户的项目评分值是可行的，在数据集比较稀疏时，基于线性回归的算法预测效果好一些(其 MAE 值比较小)，

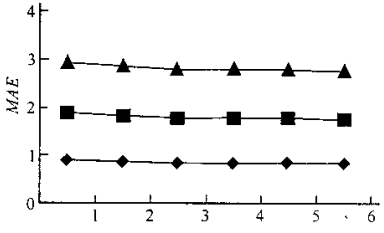

图 1 相似度预测算法的精度比较
—▲—：线性回归；—■—：修正余弦；
—◆—：余弦。

(下转第 259 页)

地进行数据挖掘以改进模型，直到找出分析人员认可的有价值的结果为止。

得到最终的决策树模型之后，我们就可以使用该模型来对企业的所有客户进行分析预测，找出可能流失的客户群体的组成和特征，并根据预测结果对可能流失的客户采取恰当的营销措施以增强其忠诚度。

4　结束语

客户资源是企业的生命，保留并巩固企业的客户资源对企业来说意义重大。将数据挖掘技术应用于客户关系管理，能够帮助企业深入理解客户，得到更加准确的客户模型，从而改进营销决策和客户服务，具有十分重要的应用价值。随着数据挖掘技术的不断成熟，基于数据挖掘的分析型客户关系管理系统无疑也将获得越来越广泛的应用。

参考文献

1　Jiawei Han，Micheline Kamber. 数据挖掘概念与技术. 范　明，孟小峰译. 北京：机械工业出版社，2001.

2　Alex Berso，Stephen Smith，Kurt Thearling. 构建面向 CRM 的数据挖掘应用. 贺　奇，郑　岩，魏　藜等译. 北京：人民邮电出版社，2001.

3　Quinlan，Ross J. C4. 5：Programs for Machine Learning. San Mateo，CA：Morgan Kaufmann Publishers，1993.

（责任编辑：邓大玉　曾蔚茹）

（上接第 254 页）

但这种算法在高密度数据集上会存在过适应性，此时效果反而会差一些。

4　结束语

基于项目评分预测的主动营销算法根据项目之间的相似性初步预测用户对未评分项目的评分，在此基础上采用一种新颖的相似性度量方法计算目标用户的最近邻居。实验结果表明，本算法可以有效解决用户评分数据极端稀疏情况下传统相似性度量方法存在的问题，显著提高主动营销系统的推荐质量。模型目前还没有考虑信息量少的用户的推荐，没有考虑模式的动态更新，没有考虑专家对推荐系统的指导作用，这些问题将有待于进一步讨论。

参考文献

1　Yoon Ho Cho，Jae Kyeong Kim，Soung Hie Kim. A personalized recommender system based on Web usage mining and decision tree induction. Expert Systems with Applications，2002，23：329～342.

2　Gerard Rodiguca Mula，Hector Garcia Molina，Andreas Paepacke. Callaborative value filtering on the Web. Computer Network and ISDN Systems，1998，30：736～738.

3　Dong seop Lee，Gye Young Kim，Hyung Il Choi. A Web-based collaborative filtering system. Pattern Recognition，2003，36：519～526.

4　John S Breese，David Heckerman，Carl Kadie. Empirial analysis of predictive algorithms for collaborative filtering. Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence，1998.

（责任编辑：邓大玉　曾蔚茹）