·情报方法·

基于模糊聚类和资源平滑的协同过滤推荐*

Collaborative Filtering Recommendation Based on Fuzzy Clustering and Item Smoothing

王惠敏　聂规划

（武汉理工大学经济学院电子商务系　武汉　430070）

摘　要　在分析目前电子商务推荐系统及传统的协同过滤推荐存在问题的基础上，提出了一种新的电子商务推荐算法。该算法利用客户对商品的历史评分记录中所隐含的客户相关信息和商品相关信息来为客户推荐商品，并且将模糊聚类技术运用于商品最近邻居和客户最近邻居的查找。实验结果表明该算法能够提供更好的推荐，聚类数对推荐质量有较大的影响。

关键词　协同过滤　模糊聚类　推荐系统

1　引　言

随着 Internet 的普及，一方面，电子商务的发展正极大地改变着人们的生活，网上购物的经历让人们感受到电子商务带来的惊喜；另一方面，由于供应链和物流的发展，电子商务的规模进一步扩大，用户经常会迷失在大量的商品信息空间中，增加了用户购买所需商品的难度，用户在找到自己需要的商品之前，必须浏览大量的无关信息。为了克服商品选购的困难，电子商务推荐系统应运而生。

电子商务推荐系统是基于可得到的信息资源向客户推荐适合其需要的信息或商品的系统[1]。该系统最大的优点在于它能收集客户感兴趣的信息，并根据客户兴趣主动为客户作出个性化推荐。目前几乎所有的大型电子商务网站都不同程度地使用电子商务推荐技术。比较成功的系统有：Xerox PARC 研究中心提出的 TYPESTRY 协同过滤推荐系统；MIT 开发的用于新闻组信息推荐的 GroupLens 自动协同过滤推荐系统；Bell Core 开发的提供电影推荐的协同过滤推荐系统；Stanford 大学数字图书馆课题组开发的用于 Web 页面推荐的复合型推荐系统 FAB[2~3]。

本文讨论的协同过滤推荐技术是电子商务推荐系统中最成功的推荐技术之一，其最大优点是不需要分析对象的属性，对推荐对象没有特殊要求，能够处理非结构化的复杂对象[4]。协同过滤推荐算法主要有基于客户的协同过滤推荐（User－based collaborative filtering recommendation）和基于资源的协同过滤推荐（Item－based collaborative filtering recommendation）[5]。

基于客户的协同过滤利用客户对商品的历史评分记录，系统采用统计的方法得到具有相似兴趣爱好的邻居客户，根据邻居客户的评分数据向目标客户产生推荐。该算法的不足之处在于大部分的客户只评分了商品中的极小部分，客户间共同评分的商品很少，由此导致难以成功地定位邻居客户，影响推荐精度；另一方面在整个客户空间上计算相似客户群的过程不可避免地成为了算法的瓶颈[5]。

基于资源的协同过滤推荐依赖于商品的相似度来决定推荐，它基于这样一个假设：如果大部分客户对一些商品的评分比较相似，则当前客户对这些商品的评分也比较相似。算法的不足之处是只能推荐那些和客户当前购买的商品相类似的商品，不能挖掘客户的潜在兴趣，作出“跨类型”的推荐。

基于客户的协同过滤和基于资源的协同过滤算法均是基于客户对商品的历史评分记录进行推荐。基于客户的协同过滤算法仅将相似客户感兴趣的商品推荐给目标客户，而没有考虑到目标客户已经购买商品的相似商品也可能是该客户感兴趣的。而基于资源的协同过滤算法仅将目标客户已经购买商品的相似商品推荐给该客户，没有考虑到与其拥有共同兴趣爱好的相似客户购买的商品可能也是该客户感兴趣的。

聚类分析是数据处理的一种重要手段和工具，通过把样本按照某种相似性准则划分成各种不同的类别，从而发现人们感兴趣的内容。聚类技术已被广泛应用于大数据集的处理，目前将其应用于协同过滤可扩展性问题的改善也已经引起了研究者们的注意[6~8]。

基金项目：国家自然科学基金“基于知识网络的电子商务智能推荐系统研究”（编号：70572079）。

作者简介：王惠敏，女，1971 年生，讲师，博士，研究方向为商务智能、电子商务推荐系统、知识管理与知识工程；聂规划，男，1957 年生，教授，博导，研究方向为商务智能、信息资源管理、知识管理与知识工程。

本文的研究主要在于利用客户对商品的历史评分记录中所隐含的客户相关信息和商品相关信息来为客户推荐商品;并且通过客户对商品评分的相似性对商品进行模糊聚类和根据客户的偏好对客户进行模糊聚类,选择目标商品所在的类或目标客户所在的类作为查询空间,搜索目标商品或客户的最近邻居,从而尽量地减少计算空间;同时还着重研究聚类数的变化对推荐质量产生的影响。

2 新的协同过滤推荐算法

在电子商务交易中,有一系列的客户 $U = [u_1, u_2, \cdots, u_m]$ 和一系列的商品 $I = [i_1, i_2, \cdots, i_n]$。客户对商品的评价可由一个 $m \times n$ 阶矩阵 $R(m, n)$ 表示,第 i 行第 j 列的元素 $r_{i,j}$ 代表了第 i 个客户对第 j 个商品的评分。

2.1 平滑客户对商品的评分数据集 在实际应用中,电子商务网站的交易数据库包含的商品数量是非常巨大的,交易数据库中的客户数量也非常巨大,每个客户所关注的产品仅仅占数据库中产品种类很小的比例。因此,客户购买或评价的商品占商品总数的比例很低,造成客户对商品的评分数据集极端稀缺。由于基于客户的协同过滤和基于资源的协同过滤推荐的基础均为客户对商品的评分数据集,数据的稀疏导致很难正确找到相似客户或相似商品,推荐质量将会变差。针对数据的稀疏性问题,近期研究者们提出了许多新的方法,如基于概念分层的个性推荐方法、基于案例推理的协同过滤推荐方法等[9~10]。

本文提出的算法主要是通过采用评分数据集中隐含的商品与商品的相关性信息来降低评分数据的稀疏性。首先依据客户对商品的评分将商品进行划分,客户评分相似的商品被聚为一类,然后在该类中预测客户对未评价商品的评分,降低预测计算空间的维度。运用商品的预测评分平滑评分数据集。对商品的模糊聚类计算和未评价商品的预测评分计算比较耗时,采用离线进行处理,减少推荐时的运算时间。算法的具体步骤如下:

步骤 1:运用模糊 C-均值算法将商品划分为 c 类,相应地,评分数据集 R 被划分为 $R_1, R_2, \cdots, R_c$ 且 $R_1 \cup R_2 \cup \cdots \cup R_c = R$, $R_i \cap R_j = \emptyset$, $1 \leqslant i, j \leqslant c$, $i \neq j$。

步骤 2:查找客户 u_a 的未评分商品 i_a 所属的类,若 $i_a \in R_i$,那么 R_i 中的其他商品作为该商品的邻居。

步骤 3:在类 R_i 中采用相关相似性方法计算商品与商品的相似性,客户未评分商品 i_a 与其邻居商品 i_b 的相似度可表示为:

$$\mathrm{sim}(i_a, i_b) = \frac{\sum_{u_s \in T}(r_{u_s, i_a} - \bar{r}_{i_a})(r_{u_s, i_b} - \bar{r}_{i_b})}{\sqrt{\sum_{u_s \in T}(r_{u_s, i_a} - \bar{r}_{i_a})^2 \sum_{u_s \in T}(r_{u_s, i_b} - \bar{r}_{i_b})^2}}$$

其中,T 为所有对商品 i_a 和 i_b 已做出评价的客户的交集,$\bar{r}_{i_a}$ 为商品 i_a 的平均评分,$\bar{r}_{i_b}$ 为商品 i_b 的平均评分。

步骤 4:预测客户 u_a 对未评分商品 i_a 的评分 P,并将其预测值填充在评分数据集中。P 可表示为:

$$P = \frac{\sum_{i_b \in R_i} \mathrm{sim}(i_a, i_b) \times r_{u_a, i_b}}{\sum_{i_b \in R_i} \mathrm{sim}(i_a, i_b)}$$

对评分数据集中的所有未评分商品进行预测评分,获得资源平滑后的评分数据集 A。

2.2 具有相似偏好的客户 评分数据隐含了客户对商品的偏好,运用客户与相似客户的共同偏好来为目标客户推荐商品。本文为了降低相似客户的搜索空间,模糊 C-均值聚类算法用于查找目标客户的邻居客户。针对资源平滑后的评分数据集,运用模糊 C-均值算法将客户进行划分。每一类中的客户为具有相似偏好的客户。

2.3 产生推荐 查找目标客户 u_a 所属的类,该类中其他客户为其邻居客户。计算客户与其邻居客户的相似度 $\mathrm{sim}(u_a, u_b)$:

$$\mathrm{sim}(u_a, u_b) = \frac{\sum_{i_s \in S}(r_{u_a, i_s} - \bar{r}_{u_a})(r_{u_b, i_s} - \bar{r}_{u_b})}{\sqrt{\sum_{i_s \in S}(r_{u_a, i_s} - \bar{r}_{u_a})^2 \sum_{i_s \in S}(r_{u_b, i_s} - \bar{r}_{u_b})^2}}$$

其中,S 为初始评分数据集中两客户共同评分的商品集。

客户对商品的最后预测评分为:

$$P_{u_a, i_a} = \bar{r}_{u_a} + \frac{\sum_{u_b \in A_i}(r_{u_b, i_a} - \bar{r}_{i_a}) \times \mathrm{sim}(u_a, u_b)}{\sum_{u_b \in A_i} \mathrm{sim}(u_a, u_b)}$$

3 实验及结果分析

实验采用的数据集来自 基于 Web 的研究型推荐系统 MovieLens(http://MovieLens.umn.edu/),MovieLens 站点用于接收客户对电影的评分并提供相应的电影推荐列表。运用 MATLAB 工具将 MovieLens 数据集转换为客户—电影评分矩阵,从中随机截取 200 个客户对 800 部电影的评分数据,然后将评分数据按 0.9 的比率划分为训练集和测试集,从测试客户中的评分中随机选取 5 个评分作为可见的评分。平均绝对偏差 MAE(Mean Absolute Error)作为评价该推荐算法质量的度量标准,验证算法的有效性。平均绝对偏差 MAE 通过计算预测的客户评分与实际的客户评分之间的偏差来度量预测的准确性,MAE 越小,推荐质量越高[5]。

运用本文推荐的算法针对客户—电影评分矩阵进行模拟推荐。实验中我们首先将文中提出的推荐算法(FCIS-BCF)与传统的基于客户的协同过滤算法(UBCF)和基于资源的协同过滤算法(IBCF)相比较,结果如图 1 所示。从实验结果中可以看出,我们所提出的推荐算法具有较小的 MAE 值,在推荐质量上均明显优于基于客户的协同过滤算法和基于资源的协同过滤算法。

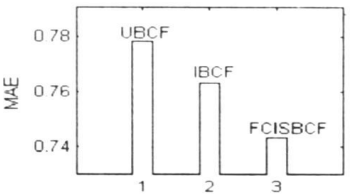

图 1　不同算法的 MAE 比较

另外，在本文推荐的算法中，我们还主要研究了聚类数对推荐质量产生的影响。首先模拟了不同商品聚类数状态下产生的推荐，结果如图 2 所示。实验结果显示了当客户聚类数一定时，商品聚类数对推荐质量的影响。从图 1 中我们可以看出，当在客户聚类数分别为 4、5、6 的状态下，商品聚类数为 12 时，MAE 值同时达到最小。

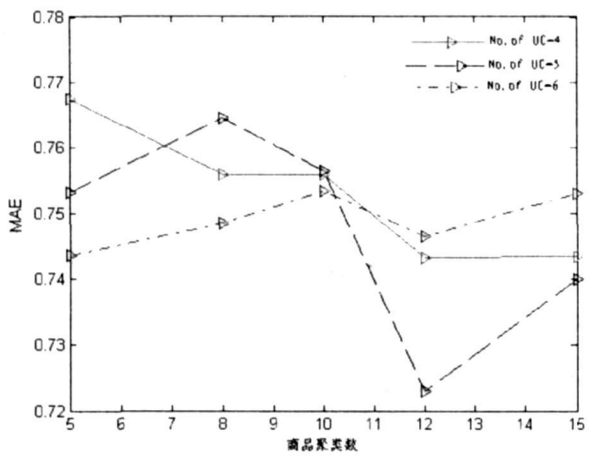

图 2　商品聚类数对 MAE 值的影响

其次，针对文中的推荐算法，本文又模拟了不同客户聚类数状态下产生的推荐，结果如图 3 所示。实验结果显示了商品聚类数一定时，客户聚类数对推荐质量的影响。由图可知，当商品聚类数分别为 5、8、10 的状态下，客户聚类数为 6 时，MAE 值同时达到最小。因此，聚类数的指定对该推荐算法的推荐质量会产生较大的影响，采用合适的聚类数不仅能有效降低预测计算的时间，提高推荐系统的实时响应速度，而且能较好地提高推荐精度。

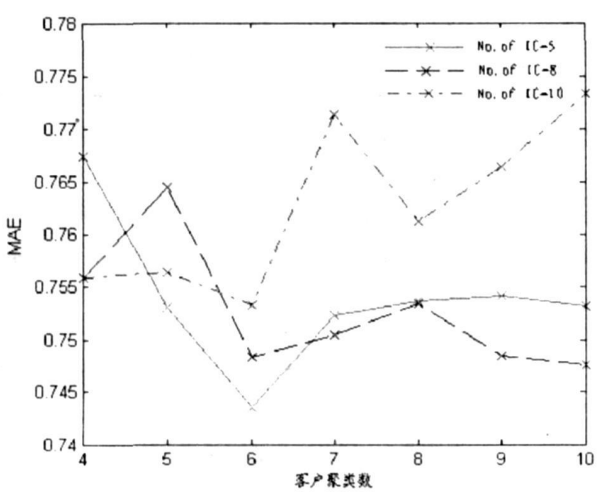

图 3　客户聚类数对 MAE 值的影响

4　结　论

在本文提出的基于模糊聚类和资源平滑的协同过滤推荐算法中，客户对商品的预测过程融合了客户对商品的历史评分记录中隐含的商品相关和客户相关的信息，降低了评分数据的稀疏性，预测结果更为准确。并且，运用模糊聚类技术将商品空间和客户空间划分为多个小的聚类，这不仅使商品最近邻居和客户最近邻居的查找能在更小的计算空间中完成，而且每个划分之间是相互独立的，其中的计算可并行进行，从而能够减少预测计算的时间。但是，聚类数的变化对推荐质量有较大的影响，采用合适的聚类数不仅能提高推荐系统的实时响应速度，而且能提高推荐精度。

参 考 文 献

1　Weng L T, Xu Y, Li Y F. An Improvement to Collaborative Filtering for Recommender Systems. Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technological and Internet Commerce, Washington: IEEE Computer Society, 2005

2　Chuan-Feng Chiu, Timothy K Shih, Ying Hong Wang. An Integrated Analysis Strategy and Mobile Agent Framework for Recommendation System in EC over Internet. Tamkang Journal of Science and Engineering, 2002, 5(3)

3　余　力, 刘　鲁, 罗掌华. 我国电子商务推荐策略的比较分析. 系统工程理论与实践, 2004, 24(8)

4　Breese J, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Madison: Morgan Kaufmann, 1998

5　Sarwar B, Karypis G, Konstan J. Item-based Collaborative Filtering Recommendation Algorithms. Proceedings of the 10th International World Wide Web Conference, New York: ACM, 2001

6　张海燕, 顾　峰, 姜丽红. 基于模糊簇的个性化推荐方法. 计算机工程, 2006, 32(12)

7　George T, Merugu S. A Scalable Collaborative Filtering Framework Based on Co-clustering. Fifth IEEE International Conference on Data Mining, Houston, USA: IEEE Press, 2005

8　Xue G R, Lin C, Yang Q, et al. Scalable Collaborative Filtering Using Cluster-based Smoothing. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM, 2005

9　熊　馨, 王卫平, 叶跃祥. 基于概念分层的个性化推荐算法. 计算机应用, 2005, 25(5)

10　席俊红. 基于案例推理: 一种用来改善稀疏性问题的方法. 微型电脑应用, 2005, 21(12)

（责编: 京梅）