

基于用户聚类的协同过滤推荐方法

查文琴, 梁昌勇, 曹 镭

(合肥工业大学 管理学院, 安徽 合肥 230009)

摘 要: 为了提高电子商务网站的个性化服务效率, 给出了一种改进的用户聚类的协同过滤推荐方法, 该算法利用用户对项目的关注的相似性来修正原始相似性计算, 综合考虑用户对项目的关注和用户评价对推荐的影响。实验表明, 该基于用户聚类的协同过滤推荐算法不仅减少了用户在寻找最近邻居的搜索强度, 加快了推荐生成速度, 而且增强了推荐算法的实时性, 提高了推荐质量。

关键词: 协同过滤; 聚类分析; 电子商务系统; 个性化推荐

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2009)06-0069-03

Collaborative Filtering Recommendation Method Based on Clustering of Users

ZHA Wen-qin, LIANG Chang-yong, CAO Lei

(School of Management of Hefei University of Technology, Hefei 230009, China)

Abstract: In order to raise service efficiency of the recommendation system, an improved collaborative filtering recommendation method based on clustering of users is proposed. This new method revises the original similarity using users' interest in item, takes synthetically into account the influence of users' interest in item and users rating. The experimental results show that the presented method not only reduces the search space for nearest neighbors but also improves the performance of CF systems in recommendation quality and efficiency.

Key words: collaborative filtering; clustering analysis; e-commerce system; personalized recommendation

0 引 言

个性化推荐是利用电子商务网站向客户提供商品信息和建议, 帮助用户决定应该购买什么产品, 模拟销售人员帮助客户完成购买过程^[1]。个性化推荐技术是电子商务推荐系统中最核心、最关键的技术, 很大程度上决定了推荐系统性能的优劣。目前个性化推荐系统中使用的主要推荐技术主要分为3类: 基于规则的推荐技术、基于内容过滤的推荐技术和最近邻协同过滤推荐技术。

迄今为止, 最近邻协同过滤是最成功的推荐技术^[2], 其基本思想是推荐系统根据目标用户与其他用户之间的相关性进行推荐, 当系统发现一个或一组用户与目标用户的消费偏好相似时, 系统就可以根据这些用户的消费行为来预测目标用户的消费行为。

最近邻协同过滤推荐需要在整个用户空间上搜索目标用户的最近邻居, 随着数据的不断增多、系统规模的扩大, 在整个用户空间上搜索目标用户的最近邻居的计算量成线性增大, 系统的实时性能越来越差, 这成为协同过滤推荐系统发展的瓶颈。针对上述问题, 提出了一种改进的基于用户聚类的协同过滤算法, 计算用户对项目的关注相似性和用户对项目的评分相似性, 并将两者线性组合, 利用组合后的相似性对用户进行聚类, 从而将具有相似兴趣的用户放入同一聚类中。当目标用户到达时, 判断用户所属聚类, 再在对应聚类中搜索目标用户的最近邻居, 从而在尽量小的用户空间上搜索目标用户的最近邻居, 最后根据最近邻居对项目的评分预测目标用户对项目评分并产生推荐列表。实验结果表明文中提出的算法在一定程度上提高了在线推荐的实时响应速度和推荐的精度。

收稿日期: 2008-09-19; 修回日期: 2008-12-04

基金项目: 国家自然科学基金重点资助课题(70631003); 国家自然科学基金资助课题(70771037)

作者简介: 查文琴(1985-), 女, 硕士研究生, 研究方向为电子商务; 梁昌勇, 教授, 博士生导师, 研究方向为决策分析和智能决策支持系统。

1 协同过滤技术的相关概念

协同过滤算法是根据基本用户对项目的评分产生对目标用户的推荐列表, 它基于这样的假设^[3]: 如果用户对一些项目的评分比较相似, 则他们对其他项目的

评分也将会比较相似。协同过滤推荐系统首先搜索目标用户的若干最近邻居,然后根据最近邻居对项目的评分预测目标用户对项目评分,从而产生推荐列表。

定义 1 推荐系统中的数据源 $D = (U, I, R)$, 其中 $U = \{User_1, User_2, \dots, User_m\}$ 是基本用户的集合; $I = \{Item_1, Item_2, \dots, Item_n\}$ 是项目集合; $m \times n$ 阶矩阵 R 是基本用户对各项目的评分矩阵, 第 i 行第 j 列的元素 R_{ij} 代表用户 i 对项目 j 的评分。

定义 2 协同过滤系统中, 计算用户的相似性可以有多种方法, 主要包括余弦相似性、Pearson 相关相似性和修正的余弦相似性 3 种^[4]。文献[2]已经通过实验对各种度量方法进行了比较, 根据其结论, 选用 Pearson 相关相似性度量较好。则用户 i 和 j 之间的相似性 $\text{sim}(i, j)$ 为:

$$\text{sim}(i, j) = \frac{\sum_{c \in I_{ij}} (R_{ic} - \bar{R}_i)(R_{jc} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{ic} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_{ij}} (R_{jc} - \bar{R}_j)^2}} \quad (1)$$

其中, R_{ic} 表示用户 i 对项目 c 的评分, \bar{R}_i 表示用户 i 对项目的平均评分, I_{ij} 表示用户 i 和 j 共同评分过的项目集合。

定义 3 已知数据源 $D = (U, I, R)$, 给定目标用户 u , 对于 $\forall i \in U$, 将 $\text{sim}(u, i)$ 最大的 l 个基本用户 i 组成集合 NS_u , 则称该集合中的元素为目标用户 u 的最近邻居。

定义 4 已知数据源 $D = (U, I, R)$, 给定目标用户 u 及其最近邻居集合 NS_u , 则用户 u 对项目 j 的预测评分 P_{uj} 为:

$$P_{uj} = \bar{R}_u + \frac{\sum_{i \in NS_u} \text{sim}(u, i) \times (R_{ij} - \bar{R}_i)}{\sum_{i \in NS_u} \text{sim}(u, i)} \quad (2)$$

其中, \bar{R}_u 和 \bar{R}_i 表示用户 u 和用户 i 对项目的平均评分。

2 基于改进的用户聚类的协同过滤算法

2.1 用于计算的用户-项目关注矩阵 A

通过对用户-项目评分矩阵 R 的整理可以得到用户-项目关注矩阵 A 。其中 $A_{ij} = 0$ 或 1, 1 表示用户 i 评分过项目 j 即用户 i 对项目 j 是关注的, 0 表示用户 i 未评分过项目 j 即用户 i 对项目 j 是不关注或未关注的。矩阵 A 如表 1 所示。

2.2 改进算法中的用户相似性计算

在文中的改进算法中需要计算用户对项目的评分相似性 $\text{sim}_r(p, q)$ 和用户对项目关注相似性 $\text{sim}_i(p, q)$, 然后将两者进行线性组合, 组合后的相似性作为用户之间的最终相似性。用户对项目的评分相似性基于用户-项目评分矩阵 R , 使用相关相似性公

式(1)来定量计算。用户对项目的关注相似性基于用户-项目关注矩阵 A 来定量计算, 设用户 p 和用户 q 在 n 维项目空间上的关注值分别看作向量 $p = \{p_1, p_2, \dots, p_n\}$, $q = \{q_1, q_2, \dots, q_n\}$, 则用户 p 和用户 q 之间的项目关注相似性为:

表 1 用户-项目关注矩阵 A

	I_1	...	I_i	...	I_n
U_1	0		1		1
...					
U_i	1		0		0
...					
U_m	0		0		0

$$\text{sim}_i(p, q) = \frac{p \cap q}{p \cup q} \quad (3)$$

然后将用户对项目的评分相似性 $\text{sim}_r(p, q)$ 和用户对项目关注相似性 $\text{sim}_i(p, q)$ 进行线性组合, 组合公式为:

$$\text{sim}(p, q) = \alpha \text{sim}_r(p, q) + (1 - \alpha) \text{sim}_i(p, q) \quad (4)$$

α 为设定的可调节的基于两种来源的用户相似性平衡因子。

2.3 改进推荐算法

根据上述的相似性度量方法, 可以对用户进行聚类, 目的是将具有较高相似性的用户归在一个聚类中, 而不同聚类中的用户兴趣爱好差别较大^[5-7]。整个算法的详细步骤如下所示:

输入: 数据源 $D = (U, I, R)$, 用户-项目关注矩阵 A

输出: k 个聚类

方法:

从项目集合 $I = \{Item_1, Item_2, \dots, Item_n\}$ 中检索所有的 n 个项目;

从基本用户集合 $U = \{User_1, User_2, \dots, User_m\}$ 中检索所有的 m 个用户;

从 m 个用户中选择评分量最多的 k 个用户作为初始聚类中心, 记为 $\{W_1, W_2, \dots, W_k\}$;

k 个聚类 C_1, C_2, \dots, C_k 均初始化为空, 记为集合 $C = \{C_1, C_2, \dots, C_k\}$;

repeat

for 每个用户 $u_i \in U$

for 每个聚类中心 $W_j \in \{W_1, W_2, \dots, W_k\}$

根据公式(4)计算 u_i 与 W_j 的相似性 $\text{sim}(u_i, W_j)$

end for

$\text{sim}(u_i, W_m) = \max\{\text{sim}(u_i, W_1), \dots, \text{sim}(u_i, W_k)\}$

聚类 $C_m = C_m + u_i$

end for

until 聚类的成员不再变化

传统的聚类算法的初始聚类中心是随机选取的, 在实验过程中发现, 聚类后会出现较多的孤立点。并且

由于协同过滤算法是在搜索最近邻居的基础上进行推荐的,所以立点进行个性化推荐。研究发现,评分量多的用户可以代表一部分用户,这些用户作为聚类中心具有很好的代表性。因此,文中选择评分量最多的 k 个用户作为初始聚类中心,经实验证明能较好地减少孤立点。

2.4 对目标用户产生推荐

根据目标用户与所属聚类中的用户之间的相似性,把相似性最高的1个用户视为目标用户的最近邻居,记为集合 NS_u ,则根据公式(2)就可以计算出目标用户对项目的预测,计算出目标用户对所有未评分项目的预测评分,然后根据评分最高的前 N 个项目推荐给目标用户,也就是目标用户的 Top- N 推荐集^[8]。

3 实验结果及其分析

3.1 实验数据

实验数据采用美国 Minnesota 大学 GroupLens 项目组提供的 MovieLens 数据集(<http://www.movie-lens.umn.edu/>)。MovieLens 是一个基于 Web 的研究型推荐系统,用于接收用户对电影的评分并提供相应的电影推荐列表。MovieLens 数据集中包含了 943 个用户对 1682 部电影的 100000 条评分数据,其中每个用户至少对 20 部电影进行了评分。

GroupLens 项目组提供的 MovieLens 数据集分成 5 个互不相交的子集,每次选择一个子集作为 test 数据集,其他四个合为一个 base 数据集,从而形成了 5 对 base 数据集和 test 数据集。文中使用 5-折交叉验证(5 fold cross validation)方法进行验证。每次选择一对 base 数据集和 test 数据集,使用 base 数据集中的记录作为基本用户,对 test 数据集中的目标用户进行推荐测试。

3.2 度量标准

评价推荐系统推荐质量的度量标准采用平均绝对偏差(Mean Absolute Error)进行度量。通过计算预测的用户评分与实际的用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^[3]。对于评分数据,预测的评分集合表示为 $\{p_1, p_2, \dots, p_n\}$,对应的实际评分集合表示为 $\{q_1, q_2, \dots, q_n\}$,则平均绝对偏差的定义为:

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (5)$$

3.3 实验结果及分析

实验1 由于 α 为设定的可调节的基于两种来源的用户相似性平衡因子,所以 α 取值可能会对推荐精

度有影响。因此在该实验中 α 取值从 0 到 1.0,每次增加 0.1,观察 MAE 的变化。实验结果如图 1 所示。

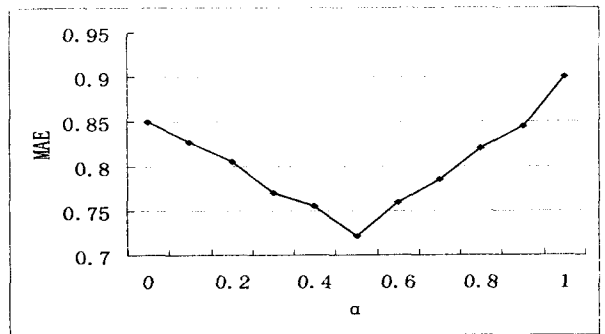


图1 平衡因子 α 对 MAE 的影响

从图1可以看出 α 取值接近 0.5 时推荐效果较好。

实验2 实验对文中基于改进的用户聚类协同过滤算法和传统的协同过滤算法的性能进行比较,根据实验1的结论 α 取 0.5。实验结果如图 2 所示。

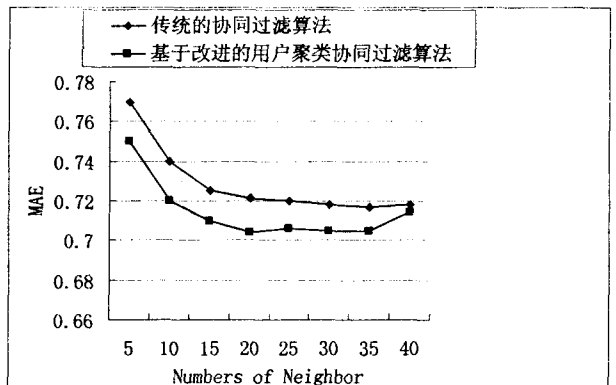


图2 与传统的协同过滤算法的比较

从实验结果的图示中可以看出,文中提出的基于改进的用户聚类协同过滤算法比传统的协同过滤算法具有较小的 MAE 值,即提出的算法在推荐精度上的确是有所改善。

4 结束语

给出的基于改进的用户聚类算法,把聚类分析用于协同过滤推荐中,既降低了用户空间的维度,又使得搜寻最近邻居用户的范围缩小,能够提高协同过滤算法的可扩展性和效率,使推荐质量比传统协同过滤要好。

参考文献:

- [1] Resnick, Varian. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.
- [2] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence

(下转第 75 页)

显示单次运行中只用了 50 秒。这已经能够满足 Q -矩阵准规则 LDPC 码这种编码方案对 Q -矩阵的要求。

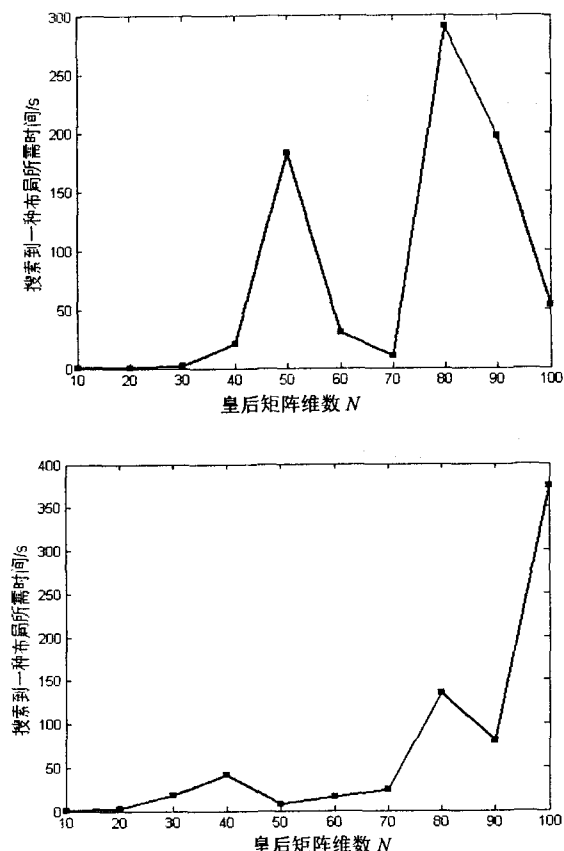


图5 随着维数 n 变化的, 单次运行所需搜索时间

4 结束语

介绍的快速搜索算法能够比经典的回溯算法更快速、灵活地找到 N 皇后问题的一种布局, 因而更满足准规则 Q -矩阵 LDPC 码这种编码方案对皇后矩阵的需求。并且在维数 N 较大时, 更加显示出其优越性。虽然该算法不能找出 N 皇后问题的所有解(能够找出所有解的算法必须是指数的), 但在实际应用中, 往往不需要得到其所有布局。

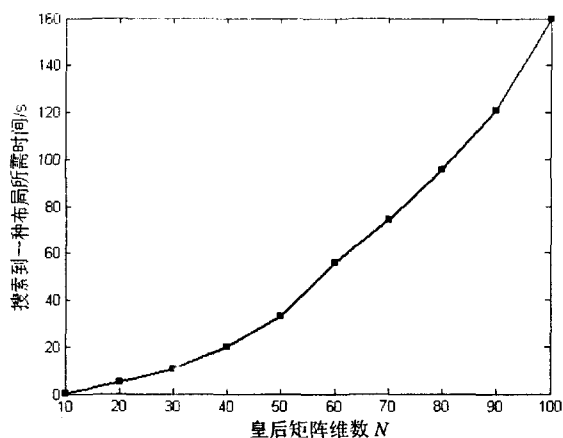


图6 随着维数 N 的变化, 循环 100 次平均搜索时间

由于此算法的随机性原理, 它的随机程度直接影响算法的性能, 算法中随机函数的随机性越强, 搜索到 Q -矩阵布局的时间就会越短。因而如果能进一步提高算法中随机函数的随机性, 此算法的性能有望进一步提高。

参考文献:

- [1] 彭立, 朱光喜. 不同置换矩阵对基于分块 H 矩阵的 LDPC 码性能的影响[J]. 计算机学报, 2008, 31(5): 783-792.
- [2] 彭立, 朱光喜. 基于 Q -矩阵的 LDPC 码编码器设计[J]. 电子学报, 2005, 33(10): 1734-1740.
- [3] 姜慧源, 田斌, 易克初. 准规则 Q 矩阵 LDPC 码编码器设计[J]. 电视技术, 2007, 31(11): 19-21.
- [4] 彭立, 朱光喜. Q -矩阵准规则 LDPC 码编码器设计的研究[J]. 计算机工程与科学, 2005, 27(6): 81-82.
- [5] 张万军. N 皇后问题回溯算法探讨[J]. 宜宾学院学报, 2006, 6(6): 64-66.
- [6] Sosic R, Gu Jun. Fast Search Algorithms for the Q -queens Problem[J]. IEEE Transactions System, Man, and Cybernetics, 1991, 21(6): 1572-1576.
- [7] Sosic R, Gu Jun. Efficient Local Search Conflict Minimization: A case Study of the n -Queens Problem[J]. IEEE Transactions on Knowledge and Data Engineering, 1994, 6(5): 661-667.

(上接第 71 页)

- (UAI-98). Sar Francisco: ACM Press, 1998: 43-52.
- [3] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceeding of the 10th International World Wide Web Conference. New York: ACM Press, 2001: 285-295.
 - [4] 余力, 刘鲁. 电子商务个性化推荐研究[J]. 计算机集成制造系统, 2004, 10(10): 1306-1312.
 - [5] 王宏超, 陈未如, 刘俊. 基于客户聚类的商品推荐方法的

研究[J]. 计算机技术与发展, 2008, 18(7): 212-214.

- [6] 程岩, 肖小云, 吴洁倩. 基于聚类分析的电子商务推荐系统[J]. 计算机工程与应用, 2005(24): 175-177.
- [7] 张娜, 何健民. 基于项目与客户聚类的协同过滤推荐方法[J]. 合肥工业大学学报: 自然科学版, 2007, 30(9): 1159-1162.
- [8] 游文, 叶水生. 电子商务推荐系统中的协同过滤推荐[J]. 计算机技术与发展, 2006(9): 70-72.