基于蚁群算法的协同过滤推荐系统的研究

吴月萍[1],王 娜[1],马 良[2]
(1. 上海第二工业大学 计算机与信息学院,上海 201209;
2. 上海理工大学 管理学院,上海 200093)

摘 要: 协同过滤算法是根据基本用户的观点产生对目标用户的推荐列表,现模拟蚂蚁觅食的原理,将用户视为具有不同属性的蚂蚁,聚类中心视为蚂蚁所要寻找的"食物源",提出基于蚁群算法实现用户聚类,以提高协同过滤推荐系统的最近邻查询速度,降低搜索开销,同时避免了使用 K-Means 聚类方法受初始聚类中心和聚类个数的影响。最终实验验证蚁群算法实现用户聚类的有效性,且解决了新用户得不到推荐的问题,并提高了协同过滤推荐算法的精确度。

关键词: 蚁群算法; 聚类; 协同过滤; 推荐; 用户
中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-629X(2011)10-0073-04

Research of Collaboration Filtering Recommendation System Based on Ant Algorithm

WU Yue-ping[1], WANG Na[1], MA Liang[2]
(1. School of Computer and Information, Shanghai Second Polytechnic University, Shanghai 201209, China;
2. College of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Collaboration filtering recommendation algorithm is that generate the recommendation list according to basic user´ view. Now imitated ant foraging theory, the users are regarded as different attributes ants, clustering center is regarded as the "food source" that the ants are looking for, proposed to cluster user based ant algorithm, for improving the query speed of the nearest neighbor in the collaborative filtering recommendation system, reducing the search spending, and avoiding the effects of initial clustering centers and clustering numbers in the use of K-Means clustering method. Finally, the experiment verify that user clustering through ant algorithm is effective, and solve the problem of new user not recommended, enhance the precision of collaboration filtering recommendation algorithm.

Key words: ant algorithm; clustering; collaboration filtering; recommendation; user

0 引 言

在电子商务推荐系统中,协同过滤推荐技术是应用最成功的个性化推荐技术之一[1]。协同过滤的基本思想是基于目标用户会对有共同爱好的邻居用户所喜欢的商品产生兴趣,采用相似性度量算法搜索有共同爱好的邻居用户,根据邻居用户所评价的项目信息,向目标用户推荐项目,然后对候选推荐项目进行预测评价产生推荐结果。随着电子商务规模的扩大[2],用户数目和商品数目呈指数级增长,传统协同过滤技术的性能越来越差。目前,文献[3~5]采用 K-Means 实现用户聚类,以解决寻找最近邻居的开销问题,而一般电子商务系统中的项目更新相对缓慢,且本身有分类体系,所以不需要对项目类别作出调整。但在此聚类方法中,初始聚类中心和聚类中心间的最小距离等参数对聚类效果的影响非常大,如何合理的选取这些参数,仍是一个需要考虑的重要问题。

计算机科学家通过模仿蚂蚁行为提出了一系列方法,且这些方法也已解决了一些问题。例:1991 年 Deneubourg 等通过蚂蚁实现聚类和分类[6]用于机器人作业调度中;2002 年 Labroche 等提出基于蚂蚁化学识别系统的聚类方法。

文中受文献[7]启发,采用基于蚂蚁觅食[8]的聚类算法。蚂蚁的觅食是一个从搜索食物到搬运食物的过程。每个蚂蚁在运动过程中通过感知信息素的强弱来选择路径,也就是说蚂蚁倾向于信息素强度高的方向移动,同时蚂蚁在其经过的路径上也释放信息素,这样经过蚂蚁越多的路径其信息素越强,整个蚁群的行为表现出信息正反馈现象。当然也要考虑信息素自身

收稿日期:2011-03-28;修回日期:2011-06-09
基金项目:国家自然科学基金资助项目(70871081);上海市重点学科建设资助项目(S30504)
作者简介:吴月萍(1979-),女,江苏常熟人,硕士,工程师,研究方向为数据挖掘、推荐算法;马 良,教授,博士,博士生导师,研究方向为算法设计、系统工程。

会随时间的流逝而挥发。

借鉴基于蚁群算法所模拟的蚂蚁觅食过程，文中将用户视为具有不同属性的蚂蚁，将用户聚类中心视为蚂蚁所要寻找的“食物源”，用户聚类过程就可以看作是蚂蚁寻找“食物源”的过程。

1 协同过滤算法相关概念

协同过滤算法是基于这样的假设：如果一些用户对某些项目的评分比较相似，则他们对其它项目的评分也将会比较相似。协同过滤推荐系统首先通过相似性度量算法搜索目标用户的若干最近邻居，然后根据相邻用户的评价产生对目标用户的推荐列表，最后预测目标用户对项目的评分，产生推荐结果。

1.1 数据源描述

推荐系统中的数据源 $D=(U,I,R)$，其中 $U=\{u_1,u_2,\cdots,u_m\}$ 是基本用户的集合，$|U|=m$；$I=\{I_1,I_2,\cdots,I_n\}$ 是项目集合，$|I|=n$；R 是 $m\times n$ 阶基本用户对各项目的评分矩阵，其中的元素 $r_{ij}\in R$ 表示用户 i 对项目 j 的评分。

1.2 相关相似性

度量用户间的相似性主要有两种方法[9]：余弦相似性和相关相似性。余弦相似性实现起来比较简单，也能较好地度量用户间的相似性，而且计算速度较快。但是在评分数据极端稀疏的情况下，通过余弦相似性寻找的邻居不够准确；相关相似性考虑了用户的平均评分，可以较好地保证寻找邻居的准确性。文献[10]分别采用余弦相似性和相关相似性进行两组实验。实验结果显示，相关相似性较余弦相似性所得的推荐质量更高。因此，文中采用相关相似性度量用户间的评分相似性，也就是说这里用户 a 和 b 之间的相似性 $\mathrm{sim}(a,b)$ 是通过 Pearson 相关系数度量的，如公式(1)。

$$\mathrm{sim}_R(a,b)=\frac{\sum_{i\in S}(R_{a,i}-\bar{R}_a)(R_{b,i}-\bar{R}_b)}{\sqrt{\sum_{i\in S}(R_{a,i}-\bar{R}_a)^2}\sqrt{\sum_{i\in S}(R_{b,i}-\bar{R}_b)^2}} \tag{1}$$

其中，S 表示用户 a 和 b 共同评分的项目集合，$R_{a,i}$ 和 $R_{b,i}$ 分别表示用户 a 和 b 对项目 i 的评分，$\overline{R_a}$ 和 $\overline{R_b}$ 分别表示用户 a 和用户 b 对项目的平均评分。

1.3 预测评分

将相似度最高的若干用户作为目标用户 u 的邻居集合 NS_u，其中 $u\notin NS_u$，且集合 NS_u 中的用户按照与 u 的相似度从高到低排列。根据相似邻居预测用户 u 对未评分项目 i 的评分 $P_{u,i}$ 为：

$$P_{u,i}=\overline{R_u}+\frac{\sum_{a\in NU_u}\mathrm{sim}(u,a)\times(R_{a,i}-\bar{R}_a)}{\sum_{a\in NU_u}|\mathrm{sim}(u,a)|} \tag{2}$$

其中，$\overline{R_u}$ 和 $\overline{R_a}$ 分别表示用户 u 和用户 a 对项目的平均评分。

2 基于蚁群算法的协同过滤推荐系统

文中基于蚁群觅食原理，实现协同过滤算法的用户聚类，以提高最近邻的查询速度，且解决了使用 K-Means 聚类需人工确定 k 个类及聚类中心的问题，并基于用户属性获得能见度来解决新用户冷启动问题，避免新用户永远得不到推荐的情况。

2.1 数据预处理

数据预处理是数据优化、格式转化的过程，通过此过程建立用户模型。文中使用空间向量模型表示用户，能够便于后面的用户聚类和相似性计算。预处理过程具体包括数据清理、数据转化、归一化处理等。

(1) 数据清理具体来说是一个数据优化的过程。删除那些不符合要求的描述信息、不必要的用户属性以及错码、乱码等，使更加有效地获取高质量的用户分类。

(2) 数据转化是将用户的属性值用向量来表示。用户属性信息具有一定范围的属性值，分类较稳定，一般很少变化，容易维护。

(3) 归一化处理是把属性值限制在需要的一定范围内。首先归一化是为了后面数据处理的方便，其次是保证程序运行时收敛加快。

2.2 蚁群聚类算法

蚂蚁在觅食活动中能够在它所经过的路径上释放信息素，而且能够感知信息素的存在及其强度，并以此指导自己的运动方向。借鉴这一原理，将用户数据视为具有不同属性的蚂蚁，用户聚类中心看作是蚂蚁所要寻找的“食物源”，所以用户聚类过程就看作是蚂蚁寻找“食物源”的过程。

输入： 数据源 $D=(U,I,R)$，参数 α 和 β

输出： 用户聚类

(1) 设 U 是 m 个 l 维待进行聚类分析的用户数据集合，$U=\{U_i\mid U_i=(U_{i1},U_{i2},\cdots,U_{il}),i=1,2,\cdots,m\}$，其中 U_{ij} 指用户 i 的第 j 个特征属性值，则数据对象间的欧氏距离(相似度)为：

$$d(u_i,u_j)=\sqrt{\sum_{r=1}^{l}(u_{ir}-u_{jr})^2} \tag{3}$$

u_i,u_j 特征属性越相似，其欧氏距离越小，反之越大。

(2) 根据数据源 $D=(U,I,R)$，运用公式(1)，可以获得用户的评分数据相似度 $\mathrm{sim}(u_i,u_j)$，考虑到项目评分的不同时间会影响用户间的评分相似度，故最终相似度 $\mathrm{sim}(u_i,u_j)\times w_{ij}$，式中权值 $w_{ij}=1/||t_i-t_j||$，t_i 和 t_j 分别为用户 u_i 和 u_j 进行项目评分时的时间

段,时间间隔越短,w_{ij} 值越大,反之越小。若用户评分为同一时间段,则设 w_{ij} 为 1。最后以附带权值的评分相似度作为蚁群算法中的信息素 τ_{ij}。

(3) 运用蚂蚁觅食原理,蚂蚁寻找食物源所选择路径的概率(4) 作为判断用户 u_i 是否与 u_j 归为一类的依据。

$$p_{ij} = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{s \in S}\tau_{sj}^{\alpha}\eta_{sj}^{\beta}} \tag{4}$$

式中,$S = \{u_s \mid s \neq j, s = 1,2,\cdots,m\}$,$\eta_{ij} = 1/d(u_i, u_j)$ 称为能见度,这个量基本不变,α 和 β 为控制信息素和能见度之间的可调节参数。当参数 $\alpha = 0, \beta = 1$,以用户能见度来寻找聚类集,因此,即使那些从未评分过的新用户,也能通过蚁群算法找到聚类集,从而解决了新用户得不到推荐的问题。

重复步骤(1) – (3),计算 $p_{ij+1}, p_{ij+2}, \cdots, p_{im}$,寻找与 u_i 最大的 $\max(p_{ix})$,$x = 1,2,\cdots,m, x \neq i$,则将 u_i 归并到 u_x 领域。这里没有涉及 P_{ix} 与设定概率 P_0 的比较,不管最大概率是否大于 P_0,文中都以与 u_i 的最大概率的 u_x 作为归并类,这样解决了用户聚类的孤立点问题。

2.3　协同过滤推荐算法

基于蚁群聚类算法获得用户数据对象的聚类,然后在聚类中寻找目标用户 u_i 的最近邻居,基于最近邻居评分实现目标用户向未评分项目的预测评分,并从中选择前 N 个评分最高的项目,作为目标用户 u_i 的 Top – N 推荐集。

输入: 数据源 $D = (U, I, R)$,用户聚类

输出: 最高评分的 N 个推荐集。

(1) 根据公式(1) 计算目标用户 u_i 在此聚类中的用户评分相似度,并按从高到低排列形成邻居集合 $M = \{u_1, u_2, \cdots, u_t\}$,邻居集中邻居个数 t 可通过实验设定。

(2) 根据公式(2) 及最近邻居集 M 即可预测目标用户 u_i 对未评分项目的评分,最后选择前 N 个评分最高的项目推荐给目标用户。

3　实验结果与分析

3.1　实验环境

以 MovieLens 站点所提供的数据集为实验环境,其中用户 943 个,项目(影片) 1682 个,用户对影片产生 10 万条评分记录,但用户评分数据集的稀疏等级只有 0.9370。

每个用户由编号、年龄、性别、职业等属性描述,将各用户年龄及性别属性值分别转化成向量表示,而职业属性不能独立转化,需进行用户间的比较,若用户职业相同,则此项比值为 0,否则比值为 1,经过这些数据处理后,计算用户间的欧氏距离。用户对影片按五个(1,2,3,4,5) 等级来评定,整个数据集按 80% 和 20% 来进行划分成训练集和测试集。

3.2　评价指标

评价推荐系统推荐质量的度量标准有统计精度度量方法,该方法中的平均绝对偏差 MAE(Mean Absolute Error) 易于理解,应用较直接、广泛。因此,文中采用平均绝对偏差 MAE 进行度量。通过计算用户对目标项目的预测值与实际评价值之间的偏差来度量评价预测的准确性。其偏差越小,预测精度越高,推荐质量越高,否则相反。平均绝对偏差 MAE 定义为[11]:

$$\mathrm{MAE}_u = \frac{\sum_{i=1}^{N} |p_{u,i} - q_{u,i}|}{N} \tag{10}$$

其中,$p_{u,i}$ 表示用户 u 对项目 i 的预测评分,$q_{u,i}$ 表示用户 u 对项目 i 的实际评分,N 为被评估的用户 u 在测试集中待评价的影片个数。

3.3　实验结果与分析

实验 1 基于蚁群聚类算法实现 MovieLens 站点中的用户集分类。实验首先比较带时间差和不带时间差的评分相似度,结果显示: 考虑评分时间因素的 943 个用户,在参数 α 和 β 的调节过程中,其聚类数都要达到二百多个,可见聚类个数较多,分类较细; 而不考虑时间差的聚类,其聚类个数明显减少,当参数 $\alpha = 0, \beta = 1$ 时,聚类个数为 13,当参数 $\alpha = 1, \beta = 0$ 时,聚类个数为 21,其余参数值,所得聚类数均为一百多个。因此,虽然考虑时间因素的评分能更精确地实现用户聚类,但对于评分数据相当稀疏、聚类数据量不大的环境,可以不考虑时间差。文中实验 2 未考虑时间差,当目标用户为新用户时,即没有评分记录的用户,则设参数 $\alpha = 0, \beta = 1$,以用户能见度来寻找聚类集; 否则使用参数 $\alpha = 1, \beta = 0$,凭用户信息素来实现聚类。

实验 2 基于实验 1 所得的用户聚类,对已有用户评分的项目进行预测评分,分别取 10 至 50 个最近邻居数进行预测评分,间隔为 10。在同一数据环境下与传统协同过滤(Traditional CF)、基于 K–Means 用户聚类的协同过滤进行比较。最终结果如图 1 所示,基于蚁群算法实现协同过滤推荐能得到相对较好的效果。

4　结束语

为了避免受 K–Means 聚类初始值设置的较大影响,文中基于蚂蚁觅食的原理,通过蚂蚁寻找“食物源”的过程实现用户聚类,实验验证此方法的在降低最近邻搜索开销的同时,提高了协同过滤推荐的精度,且调节参数 α 和 β,可以解决新用户的冷启动问题。

针对用户评分相对密集的数据，还可以考虑时间差的因素，以进一步提高推荐效果；同时可以考虑文献[12]的方法实现用户属性的加权，以提高聚类的准确性。

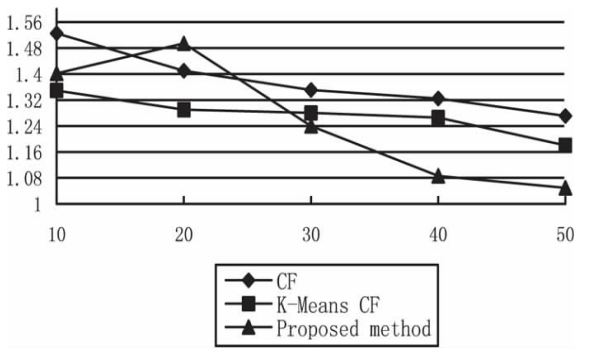

图1 推荐精度的比较

参考文献：

[1] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering [C]//In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). [s. l.]: [s. n.], 1998: 43-52.

[2] Lee J S, Jun C H, Lee J, et al. Classification-based collaborative filtering using market basket data [J]. Expert System with Applications, 2005, 29(3): 700-704.

[3] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[4] 李 涛，王建东. 一种基于用户聚类的协同过滤推荐算法[J]. 系统工程与电子技术，2007，29(7)：1178-1182.

[5] 黄国言，李有超. 基于项目属性的用户聚类协同过滤推荐算法[J]. 计算机工程与设计，2010，31(5)：1038-1041.

[6] Deneubourg J L, Goss S, Franks N, et al. The dynamics of collective sorting: Robot-like ants and ant-like robots [C]//Proceedings of the First international Conference on Simulation of Adaptive haviour: From Animals to Animals J. Cambridge, MA: MIT Press, 1991: 356-365.

[7] 杨 燕，张昭涛. 基于阈值和蚁群算法结合的聚类方法[J]. 西南交通大学学报，2006，41(6)：719-742.

[8] 马 良，朱 刚，宁爱兵. 蚁群优化算法[M]. 北京：科学出版社，2008.

[9] Aggarwal C C. On the effects of dimensionality reduction on high dimensional similarity search [C]//Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART. Symposium on Principles of Database Systems. [s. l.]: [s. n.], 2001: 256-266.

[10] 王明文，陶红亮. 双向聚类迭代的协同过滤推荐算法[J]. 中文信息学报，2008，7(22)：61-65.

[11] Sarwar B, Karypis G, Konstan J, et al. Item-Based collaborative filtering recommendation algorithms [C]//In: Proceedings of the 10th International World Wide Web Conference. [s. l.]: [s. n.], 2001: 285-295.

[12] 李玲娟，李 冰. 一种基于特征加权的蚁群聚类新算法[J]. 计算机技术与发展，2010，20(8)：67-70.

(上接第72页)

5 结束语

文中将不确定数据中形成的可能世界进行了缩减，以此为基础进行k个最佳结果查询。通过实验表明文中提出的RPW-KBest算法显著地提高了查询效率，减小内存消耗。由于概率查询算法依然面临很多挑战和尚待解决的问题，因此，下一步的工作综合分析现有查询算法的优缺点，在查询算法上进行深入研究。

参考文献：

[1] 周傲英，金澈清，王国仁，等. 不确定性数据库管理技术研究综述[J]. 计算机学报，2009，32(1)：1-16.

[2] Soliman M A, Ilyas I F, Chang KevinChen-Chuan. Top-k Query Processing in Uncertain Databases [C]//2007 IEEE 23rd International Conference on Data Engineering. [s. l.]: [s. n.], 2007: 15-20.

[3] 孙永佼，王国仁. P2P环境中不确定数据Top-k查询处理算法[J]. 计算机研究与发展，2009，46(S)：280-286.

[4] R'e C, Dalvi N, Suciu D. Efficient Top-k Query Evaluation on Probabilistic Data [C]//IEEE 23rd International Conference on Data Engineering. [s. l.]: [s. n.], 2007: 15-20.

[5] Lian Xiang, Chen Lei. Top-k Dominating Queries in Uncertain Database [C]//Data Engineering. ICDE 2007. IEEE 23rd International Conference. [s. l.]: [s. n.], 2007.

[6] Pei J, Jiang B, Lin X, et al. Probalitistic kyline on uncertain data [C]//Proceeding of the 33rd international conference on very large databases. Vienna, Austria: [s. n.], 2007.

[7] Huang Y, Chen C, LEEC. Continuous k-nearest neighbor query for moving objects with uncertain velocity [J]. Geoinformatica, 2009, 13(1): 1-25.

[8] 周 帆，李树全，肖春静. 不确定数据Top-k查询算法[J]. 电子测量与仪器学报，2010，30(10)：2605-2609.

[9] 韩希先，杨东华，李建中. TKEP：海量数据上一种有效的Top-K查询处理算法[J]. 计算机学报，2010，33(8)：1405-1418.

[10] 周 逊，李建中，石胜飞. 不确定数据上两种查询的分布式聚集算法[J]. 计算机研究与发展，2010，47(5)：762-771.

[11] 刘德喜，万常选，刘喜平. 不确定数据库中基于x-tuple的高效Top-k查询处理算法[C]//第26届中国数据库学术会议论文集(A辑). [出版地不详]: [出版者不详]，2009：15-18.

[12] 吴义虎，武志平. 基于平均车速和车速标准差的路段安全方法[J]. 公路交通科技，2008，25(3)：139-142.