文章编号: 1008 - 7133(2013)05 - 0090 - 05

基于共同评分的协同过滤算法

郭迎迎, 王 波, 周继平
(上海理工大学 管理学院, 上海 200093)

摘 要: 协同过滤是目前电子商务推荐系统中使用最广泛最成功的一种个性化推荐算法。受数据稀疏性影响,传统协同过滤算法在较小共同评分项集上计算出的相似度不能准确反映用户间的相似关系,严重影响了推荐系统的精度。针对该问题,在分析共同评分分布及其与相似度关系的基础上,提出了基于共同评分的协同过滤算法,无须计算相似度,直接将共同评分作为最近邻选择标准。MovieLens 实验表明该算法能明显提高预测结果的准确性和覆盖率。

关 键 词: 电子商务; 协同过滤; 共同评分
中图分类号: TP 311 **文献标志码:** A

The collaborative filtering based on co-ratings

GUO Ying-ying, WANG Bo, ZHOU Ji-ping

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Collaborative filtering is one of the most extensive and successful personalized recommendation algorithm in e-commerce recommendation system. Affected by data sparsity, the traditional collaborative filtering algorithms does not reflect the interest similarity of uses calculating similarity between users on the smaller set of common rated items accurately, seriously affecting the accuracy of recommendation system. To solve this problem, collaborative filtering algorithm based on co-ratings was proposed by analyzing the distribution of co-ratings and relationship between co-ratings and similarity, directly using co-ratings as a criterion to select nearest neighbor without calculating similarity. Experiments on MovieLens datasets show that the algorithm can make a substantial increase in prediction accuracy and recommendation coverage.

Key words: e - commerce; collaborative filtering; co-ratings

随着互联网的飞速发展和信息化水平的快速提高,我们从信息贫乏过渡到信息过载。各大电子商务网站,比如亚马逊、淘宝、京东、当当等,均使用各种形式的推荐系统为客户提供商品信息和购买建议。电子商务推荐系统作为一种新兴技术,能帮助顾客从海量商品信息中找到自己喜欢的商品,一方面增加网站的智能性,另一方面提高网站交叉销售的能力,增加客户对网站的忠诚度和黏性[1]。协同过滤算法是电子商务推荐系统中应用最广泛最成功的一种个性化推荐算法,基本思想是依据用户行为(评分、浏览、购买)获得其兴趣和喜好,推荐满足用户需求的商品。

在基于协同过滤的推荐系统中,K 近邻模型是使用最广泛的经典模型[2],主要原理是利用评分相似度构造 K 个最近邻居,然后根据最近邻对用户进行推荐。传统协同过滤算法常常使用 Pearson 相关系数[2],在共同评分项目的基础上计算相似度,然而由于数据稀疏,用户共同评分的项目较少,导致所计算出的相似度不能正确反映用户间的兴趣相似关

收稿日期: 2013 - 07 - 29
基金项目: 上海市重点学科基金项目(S30504,S30501)
作者简介: 郭迎迎(1988—),女,硕士研究生;
王 波(1960—),男,教授,博士,硕士生导师;
周继平(1988—),男,硕士研究生.

系。针对该问题,本文提出了基于共同评分的协同过滤算法,无需计算相似度,直接将共同评分的大小作为最近邻的选择标准和预测评分时的权重。实验结果表明该算法能大幅提高预测准确性和推荐覆盖率,改善推荐系统的精度。

1　相关工作

为了提高协同过滤算法的准确性,国内外许多学者从相似度和邻居选择的角度,提出了许多改进算法。相似度方面如 Herlocker 等[3] 最早提出了 Max 权重相似度,使用参数调节基于较少共同评分所产生的过高相似度值,并通过实验证明引入参数能够提高预测的准确性。然而当用户间的共同评分数大于参数阈值时,调整后的相似度值会大于 1,为此,Mclaughlin 等[4] 提出了 Min 权重相似度,然而 Max 和 Min 权重会根据参数分段,不够平滑,为此 Bell 等[5] 提出了 Shrink 权重相似度,采用连续平滑的方式修正相似度值。由于这些权重相似度值都引入了依赖数据集的参数,需要交叉验证才能获得最优值,使用很不方便,针对该问题,张迎峰等[6] 提出了修正重叠度因子,使用共同评分和方差修正相似度。邻居选择方面如张尧等[7] 综合考虑邻居选择权重,提出了基于用户分类的邻居选择方法。罗辛等[8] 提出相似度支持度的概念并将其作为一种新的 K 近邻度量。贾冬燕等[9] 根据相似度和信任度,提出了一种双重邻居选择策略。Huete 等[10] 根据邻居预测过去评分准确性的能力选择最近邻居。

上述某些方法虽然考虑了共同评分对相似度的影响,但是相似度仍然是主要的邻居选择标准,数据稀疏条件下由较少共同评分所得到的相似度不能准确地反映用户间的相似性,故上述方法均存在不足之处。为此本文提出了基于共同评分的协同过滤算法,直接使用共同评分作为近邻选择标准,无需计算相似度。

2　传统协同过滤算法

协同过滤算法一般分为 3 步: ①数据表述。给定用户集 U 和项目集 I,则用户对于项目的兴趣可以表示为一个 $m \times n$ 的矩阵 R。在该矩阵中,m 代表用户数,每一个行向量表示特定用户的评分集合,n 代表项目数,每一个列向量表示特定项目的被评分集合,每一元素 $r_{ui} \in \boldsymbol{R}$ 表示用户 u 对项目 i 的评分。②近邻选择。按照相似度从大到小为当前用户或项目选择最近邻集合。③推荐产生。利用最近邻居评分的加权平均值预测目标用户未评分项目的评分,根据预测评分值可以得到用户对任意项目的兴趣度及其推荐集。

2.1 计算相似度

我们采用两种主要基于共同评分的相似度: 相关相似性(Pearson) [4] 和 Jaccard 相似性[11] 。假设 I_{ij} 表示用户 u_i 和 u_j 共同评分的项目集合,那么用户 u_i 和用户 u_j 的 Pearson 相似度和 Jaccard 相似度可以分别由式(1) 和式(2) 得到

$$\mathrm{sim}(u_i,u_j) = \frac{\sum_{c\in I_{ij}} (R_{ic}-\overline{R_i})(R_{jc}-\overline{R_j})}{\sqrt{\sum_{c\in I_{ij}} (R_{ic}-\overline{R_i})^2}\sqrt{\sum_{c\in I_{ij}} (R_{jc}-\overline{R_j})^2}}; \tag{1}$$

$$\mathrm{sim}(u_i,u_j) = \frac{|I_i \cap I_j|}{|I_i \cup I_j|}。 \tag{2}$$

其中 R_{ic}和 R_{jc}分别表示用户 u_i 和用户 u_j 对项目 c 的评分,$\overline{R_i}$和$\overline{R_j}$分别表示用户 u_i 和用户 u_j 对已评分项目的平均评分,I_i,I_j 为用户 u_i 和 u_j 的已评分项目集合。

2.2　选择近邻

邻居选择通常采用 K 近邻策略,即选择与当前用户相似度最高的前 k 个用户作为邻居。对于一个活动用户 a,要产生一个依相似度由大到小排列的邻居集合 $U=\{u_1,u_2,\cdots,u_k\}$,$a \notin U$。

2.3　产生推荐

利用最近邻居预测目标用户对未评分项目的评分,从而形成 TOP-N 推荐,目标用户 u_t 对项目 i 的评分 $P_{u_t,i}$预测为

$$P_{u_t,i} = \overline{R_{u_t}} + \frac{\sum_{v\in N_{u_t}} \mathrm{sim}(u_t,v) \times (R_{v,i}-\overline{R_v})}{\sum_{v\in N_{u_t}} (|\mathrm{sim}(u_t,v)|)}。 \tag{3}$$

其中: N_{u_t}为用户 u_t 的最近邻居集合,$\mathrm{sim}(u_t,v)$ 为用户 u_t 和用户 v 的相似度,$\overline{R_{u_t}}$为用户 u_t 的平均评分,$\overline{R_v}$为用户 v 的平均评分,$R_{v,i}$为用户 v 对项目 i 的评分。

2.4　传统相似度计算方法分析

随着电子商务的快速发展,电子商务推荐系统的规模快速增长,用户不仅在项目空间上的评分变得稀疏,而且用户之间的共同评分变得更少。研究表明大型电子商务系统中,用户评分项目数不超过项目总数的 1% [12] 。因此数据稀疏条件下,传统协同过滤算法中基于共同评分的相似度有很多弊端。

1) 如果用户间共同评分的项目数为 1,那么 Pearson 相似度计算公式的分子和分母相同,相似度值为 1。

2) 如果用户间共同评分的项目数较少，那么根据仅有的几个共同评分所计算出的相似度是不准确的。

3) 如果用户对某个项目的评分和用户的平均评分相同，那么 Pearson 相似度计算公式的分母为0，相似度将无法计算。

通过上面的分析可知，稀疏数据条件下，协同过滤算法仅在较少共同评分项目上计算的相似度不能准确反映用户间的兴趣相似性，降低了目标用户最近邻的有效性和准确性，从而使得推荐系统质量大大降低。

3 基于共同评分的协同过滤算法

人和人产生认同感，往往取决于两者之间有没有或者有多少共同经历，比如看过同一部电影，喜欢听同一歌手的歌，看同一 NBA 球队的比赛等。现实生活中，更容易和彼此认同的人成为朋友，接受他们的建议，比如当要了解一部电影是否值得观看时，会向身边曾经一起看过相同电影的朋友寻求帮助，而不会去询问一个从来都没有和我们看过相同电影的人。由此，本文将用户间共同评分项目数作为最近邻居的选择标准，两者共同评分项目数越多，表明2个用户看过的电影越多，两者的兴趣相似的可能性就越大。与传统协同过滤算法类似，基于共同评分的协同过滤算法也分为 3 步骤: 计算共同评分项目数、选择邻居、产生推荐。

3.1 计算共同评分项目数

共同评分项目数为 2 个用户已评分项目交集的大小。

$$c(u,v) = |I_u \cap I_v|。 \tag{4}$$

其中: I_u，I_v 为用户 u 和 v 已评分项目集合。

3.2 选择近邻

邻居选择策略同样采用 K 近邻策略，即选择与当前用户 a 共同评分最多的前 k 个用户作为 a 的最近邻居集 $U=\{u_1,u_2,\cdots,u_k\}$，$a \notin U$。

3.3 产生推荐

选择最近邻居后，根据最近邻协同过滤的思想，利用邻居来为目标用户进行推荐，使用的预测评分公式为

$$P_{u_t,i} = \overline{R_{u_t}} + \frac{\sum_{v \in N_{u_t}} c(u_t,v) \times (R_{v,i} - \overline{R_v})}{\sum_{v \in N_{u_t}} c(u_t,v)}。 \tag{5}$$

尽管 $c(u,v)$ 可能会大于 1，但其作为归一化因子，预测评分不会受其影响。

4 实验结果及分析

4.1 数据集

本文使用 MovieLens 数据集来测试改进后的算法，该数据集是由 GroupLens 研究产品组(http://www.grouplens.org) 提供的一个著名电影评分数据集，包含 943 个用户对 1682 个电影的 10 万条打分记录，且每个用户至少对 20 部电影进行过评分，本实验从 10 万条记录中随机选取 80% 作为训练集，剩下的 20% 作为测试集。

4.2 算法评价标准

1) 平均绝对误差。平均绝对误差(Mean Absolute Error，MAE) 通过计算预测的用户评分与实际用户评分之间的偏差来度量预测的准确性[12]。MAE 通常定义为

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}, \tag{6}$$

其中: 预测的用户评分集合为$\{p_1,p_2,\cdots,p_n\}$，相应的实际用户评分集合为$\{q_1,q_2,\cdots,q_n\}$。显然，MAE 越小，算法准确性越好。

2) 覆盖率。覆盖率(*Coverage*) 主要用于衡量推荐算法所能做出预测的项目百分比[12]。用户 u 的覆盖率为

$$C_u = \frac{K_u}{N_u}, \tag{7}$$

其中: K_u 为预测结果集 $P_u=\{p_{u1},p_{u2},\cdots,p_{uj}\}$ 不为空的项目数，N_u 为需要预测的项目数，可见 Coverage 值越大，效果越好。

4.3 实验结果及分析

4.3.1 共同评分分布

分析用户间共同评分分布，如图 1 所示。

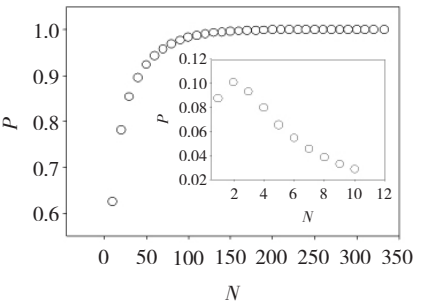

图1 共同评分数的分布

横坐标 N 为共同评分数，纵坐标 P 为共同评分数占所有评分数的累积比率，可以看到大部分用户的共同评分数都没有超过 50，只有不到 8% 的共同

评分数是在 50 以上的,然而共同评分数小于 10 的比率高达到 62.4%,具体分布情况见子图,子图纵坐标 P 为共同评分数占所有评分数的比值,从子图中可以看到用户项目评分矩阵中有接近 30% 的共同评分数是小于等于 3 的,这也就是说有 30% 的用户相似度是在不超过 3 个共同评分上计算的,可想这样得到的相似度的可信度是很低的,不能准确反映用户间的兴趣相似度。

4.3.2　Pearson 相似度与共同评分之间的关系

对 Pearson 相似度及其相应的共同评分进行分析,实验结果,如图 2 所示。

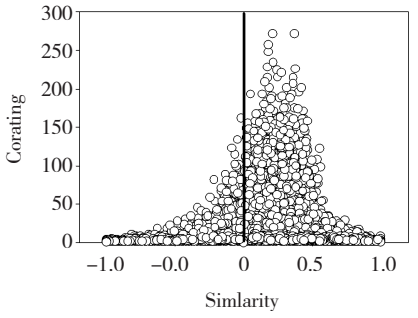

图 2　相似度值与共同评分之间的关系

散点图横坐标 Similarity 为用户间的相似度,纵坐标 Corating 为用户间的共同评分数。从图中可以看出,整个散点图向右倾斜,说明用户间共同评分越多,相似度就越高,兴趣就越相似。散点图从下到上由密集到稀疏,说明较低相似度值通常是在较少的共同评分上计算的,特别的是散点图下面的点向两端发散,说明在 −1 和 1 附近的相似度虽然较高,但通常是由较少共同评分计算而来的,因此这些相似度值的可信度很低。

稀疏数据下 K 近邻协同过滤算法根据相似度选择邻居,使得一些与目标用户共同评分非常少而相似度又较高的用户成为目标用户的最近邻居参与到推荐中,大大影响了推荐的精度。而基于共同评分的协同过滤算法根据共同评分选择邻居,使得与目标用户既有较多共同评分又有较高兴趣相似度的用户成为最近邻,从而大大提高了推荐的效果。

4.3.3　准确性比较

为了检验算法的准确性,在同等数据集的基础上变换邻居个数,和基于 Pearson 相似度的协同过滤算法(Collaborative Filtering Based on Pearson, PBCF)、基于 Jaccard 相似度的协同过滤算法(Collaborative Filtering Based on Jaccard, JBCF)比较 MAE,邻居个数 k 从 10 ~ 50,间隔为 10,实验结果,如图 3 所示。

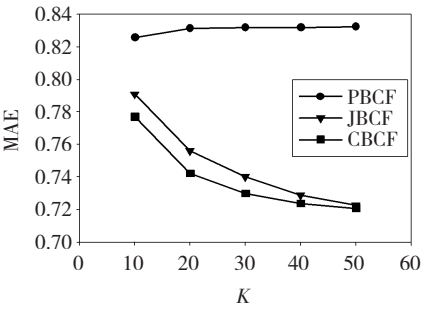

图 3　算法准确性比较

由图 3 可知,在各种实验条件下,与 PBCF 和 JBCF 相比,本文提出的基于共同评分的协同过滤算法(Collaborative Filtering Based on Co-ratings, CBCF)算法均具有较小 MAE。由此可知,本文提出的 CBCF 算法能明显提高推荐准确度。

4.3.4　覆盖率比较

为了检验算法的覆盖率,比较 CBCF 算法与 PBCF 和 JBCF 的 Coverage,邻居个数 k 从 10 ~ 50,间隔为 10,如图 4 所示。

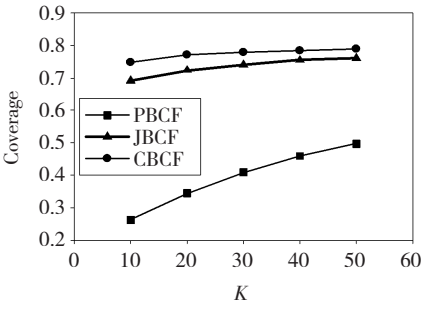

图 4　算法覆盖率比较

由图 4 可知,在各种实验条件下,无论是和 PBCF 还是 JBCF 相比,CBCF 方法都能取得较高的覆盖率。

不论是准确性还是覆盖率方面,CBCF 都优于 PBCF 和 JBCF,由此可见,共同评分能更直观准确地表示用户间的兴趣相似度,可以为当前用户选择兴趣相似的最近邻居,从而提高推荐的精度。相似度计算是协同过滤算法中比较耗时的部分,而本文提出的 CBCF 无需计算相似度,直接根据共同评分数选择最近邻居,大大改善了算法的效率。

5　结　语

稀疏数据条件下,为了克服协同过滤算法在较小共同评分上计算出的相似度不能准确反映用户间真实兴趣相似关系的问题,在深入分析了共同评分分布及其与相似度关系的基础上,提出了基于共同

评分的协同过滤算法。该算法直接根据用户间的共同评分选择最近邻，无需计算相似度，大大提高了算法的效率。MovieLens 数据集上实验表明，CBCF 能够获得比传统协同过滤算法更好地预测效果，与 PBCF 和 JBCF 相比，MAE 分别提高了 11.04% 和 1.19%，与 JBCF 相比，Coverage 提高了 5.19%.

CBCF 算法基于用户项目评分矩阵，仅利用了用户间的共同评分数，未考虑评分相似度，下一步将研究如何进一步地结合评分相似度，更好地反映用户间的兴趣相似关系。

参考文献:

[1] 李聪，梁昌勇，杨善林. 电子商务协同过滤稀疏性研究：一个分类视角[J]. 管理工程学报，2011，25(1)：94－101.

[2] BREESE J S，HECKERMAN D，KADIE C. Empirical analysis of predictive algorithms for collaborative filtering [C]//Proc. of the Fourteenth Conference on Uncertainty in Artificial Intelligence，North Carolina，USA：Morgan Kaufmann Publishers Inc. 1998：43－52.

[3] HERLOCKER J，KONSTAN J A，RIEDL J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms [J]. Information Retrieval，2002，5(4)：282－310.

[4] MCLAUGHLIN M R，HERLOCKER J L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience [C]//Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，Sheffield，UK：ACM Press，2004：329－336.

[5] BELL R，KOREN Y，VOLINSKY C. Modeling relationships at multiple scales to improve accuracy of large recommender systems [C]//Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，New York，USA：ACM Press，2007：95－104.

[6] 张迎峰，陈超，俞能海. 基于最近邻用户动态重排序的协同过滤方法[J]. 小型微型计算机系统，2011，32(8)：1581－1586.

[7] 张尧，冯玉强. 协同过滤推荐中基于用户分类的邻居选择方法[J]. 计算机应用研究，2012，29(11)：4216－4219.

[8] 罗辛，欧阳元新，熊璋，等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报，2010，33(8)：1437－1445.

[9] 贾冬艳，张付志. 基于双重邻居选取策略的协同过滤推荐算法[J]. 计算机研究与发展，2013，50(5)：1076－1084.

[10] HUETE J F，FERNÁNDEZ-LUNA J M，CAMPOS L M，et al. Using past-prediction accuracy in recommender systems [J]. Information Sciences，2012，199：78－92.

[11] CANDILLIER L，MEYER F，FESSANT F. Designing specific weighted similarity measures to improve collaborative filtering systems [C]//Proc. of the Industrial Conference on Data Mining，Berlin，Germany：Springer Verlag，2008，50(77)：242－255.

[12] RESNICK P，IACOVOU N，SUCHAK M，et al. GroupLens：an open architecture for collaborative filtering of net news [C]//Proc. of the 1994 ACM Conference on Computer Supported Cooperative Work，New York，UAS：ACM Press，1994：175－186.

[13] 刘建国，周涛，郭强，等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学，2009，6(3)：1－10.

[编辑：王劲松]