第 35 卷 第 22 期 计 算 机 工 程 2009 年 11 月
Vol.35 No.22 Computer Engineering November 2009
·软件技术与数据库· 文章编号：1000—3428(2009)22—0053—03 文献标识码：A 中图分类号：TP311.52

基于相关均值的协同过滤推荐算法

陈志敏，沈 洁，赵 耀
(扬州大学信息工程学院，扬州 225009)

摘 要：针对在用户评分数据极端稀疏环境下传统协同过滤推荐算法存在的弊端，从提高邻居用户识别准确性出发，对传统相似性度量方法进行改进，在此基础上提出一种基于相关均值的推荐算法。实验结果表明，该算法能增强邻居用户在推荐中的影响力，有效提高推荐精度，改善推荐质量。

关键词：协同过滤；相似性度量；相关均值；平均绝对偏差

Collaborative Filter Recommendation Algorithm Based on Correlation Mean

CHEN Zhi-min, SHEN Jie, ZHAO Yao
(Institute of Information Engineering, Yangzhou University, Yangzhou 225009)

【Abstract】According to the disadvantage of the traditional collaborative algorithm while the user rating data extremely sparse, this paper proposes a novel similarity measure method and a recommendation algorithm based on Correlation Mean(CM). Experimental results show it can enhance the neighbor's influence in the course of recommendation, and improve the accuracy and the quality of recommendation system effectively.

【Key words】collaborative filter; similarity measure; Correlation Mean(CM); Mean Absolute Error(MAE)

1 概述

随着互联网的普及，推荐系统已经成为电子商务环境下一种重要的个性化服务形式，通过研究目标用户的兴趣，主动为用户推荐最需要的资源，有效提高了用户面对庞大网络资源时的搜索效率。目前几乎所有大型的电子商务系统(如 Amazon、eBay 等)都不同程度地使用了各种形式的推荐系统。

协同过滤是迄今为止最成功的一种个性化推荐技术[1]。其基本思想是根据不同用户兴趣的相似性来推荐资源(项目)。核心工作是基于用户对项目的评分度量用户间的相似性，寻找与目标用户兴趣最相似的邻居用户，并采取一定算法将邻居用户评分较高的项目推荐给目标用户。随着系统中项目数量的日益增加，用户所评分项目占项目总数的比例相对变小，使整个项目空间上的评分数据极端稀疏，严重影响了邻居用户识别的准确性，导致整个系统推荐质量急剧下降。

本文提出一种新的用于度量用户相似性的加权 Pearson 相关系数方法，并在此基础上通过计算目标用户和邻居之间的相关均值对传统推荐算法进行改进。

2 相关工作

最近邻协同过滤算法是当前最典型的一种推荐技术。算法实现可以分为 3 个步骤：

(1)数据表示

在协同过滤算法中，输入数据通常表述为一个 $m\times n$ 的“用户-项目”评分矩阵 $\boldsymbol{R}_{m\times n}$，其中，$m$ 表示用户数；n 表示项目数，代表资源的类别；$R_{i,j}$ 表示第 i 个用户对第 j 个项目的评分，代表用户对该项目的喜好程度。

(2)邻居形成

准确识别目标用户的最近邻居是推荐算法的关键步骤，通常采用统计方法，通过度量用户间的相似性为需要推荐服务的目标用户寻找与其相似的 k 个最近邻居用户。其中用于相似性度量的函数主要有 Cosine 相似性和 Pearson 相关系数[2]。

(3)产生推荐

设目标用户 a 的最近邻居集合为 NBS_a，则目标用户 a 对未评分项目 i 的预测评分 $P_{a,i}$ 可以通过所有邻居用户对项目 i 评分的加权平均值进行逼近[3]，计算公式如下：

$$P_{a,i}=\bar{U}_a+\frac{\sum_{j\in NBS_a} sim(a,j)\times(R_{j,i}-\bar{R}_j)}{\sum_{j\in NBS_a}|sim(a,j)|} \tag{1}$$

其中，$\bar{U}_a$ 代表用户 a 对自身已评项目的平均评分；$sim(a,j)$ 表示用户 a 与邻居 j 之间的相似度取值；$R_{j,i}$ 表示邻居 j 对项目 i 的评分；$\bar{R}_j$ 表示邻居 j 的平均评分。采用同样方法预测用户 a 对所有未评分项目的评分，取其中评分较高的 N 个项目作为目标用户 a 的 *top-N* 推荐集。

3 基于相关均值的推荐方法

随着电子商务系统规模的扩大，用户在整个项目空间上的评分数据也相对稀疏。研究表明，在大型电子商务系统中，用户所评分项目一般不会超过项目总数的 1%[4]，这种情况下，传统推荐算法的弊端日益显露：

(1)算法在邻居形成过程中所采用的相似性度量方法存在问题。在余弦相似性方法中，对于项目空间上大量用户没

基金项目：江苏省高校自然科学基金资助项目(02KJB520013)
作者简介：陈志敏(1976－)，女，讲师，主研方向：Web 数据挖掘；沈 洁，教授；赵 耀，讲师
收稿日期：2009-03-23 **E-mail：**zmchen@yzu.edu.cn

有评分的项目，系统均将评分假设为 0，这种假设虽然可以降低计算复杂度，但并不代表实际情况。Pearson 相关系数方法虽然克服了余弦相似性度量中假设评分的负面影响，确保用户只有在共同评分项目评分相似的条件下才能取得较高相似度，但在数据稀疏情况下，即使 2 个用户共同评分项目很少，但只要两者评分相似，也能得到较高的相似性，根据常识这是不可信的。

(2)算法在产生推荐的过程中，简单基于目标用户对自身已评项目计算均值的做法，过高估计了自身评分历史对未评分项目的预测作用，削弱了邻居用户的评分对预测结果的影响，导致系统推荐质量的下降。

3.1 改进的相似度度量方法

为了解决传统相似性度量方法存在的问题，本文在 Pearson 相关系数度量方法基础上进行改进，通过引入权重因子突出用户共同评分项目规模对两者相似性的影响。改进后的相似性计算方法如下：

$$sim'(i,j)=sw\times\frac{\sum_{c\in I_{ij}}(R_{i,c}-\bar{R}_i)(R_{j,c}-\bar{R}_j)}{\sqrt{\sum_{c\in I_{ij}}(R_{i,c}-\bar{R}_i)^2}\sqrt{\sum_{c\in I_{ij}}(R_{j,c}-\bar{R}_j)^2}} \tag{2}$$

其中，$R_{i,c}$ 表示用户 i 对项目 c 的评分；$\bar{R}_i$ 和 $\bar{R}_j$ 分别表示用户 i 和用户 j 对项目的平均评分。权重因子 sw 的定义如下

$$sw=\frac{\min(|I_i\cap I_j|,D)}{D} \tag{3}$$

其中，$|I_i\cap I_j|$ 为用户 i 和用户 j 共同评分的项目数，$D(D\neq 0)$ 是算法对于用户-项目矩阵中数据稀疏状况的估计值，用于衡量数据集中用户两两间共同评价项目的平均数量。由定义可知，当两用户共同评分的项目数小于估计均值 D，即数据相对稀疏时，共同评价的项目数越多，权重越大，有效保证了只有共同评分项目较多且评分相似的用户才有可能成为邻居用户，避免了传统算法中那些共同评分项目数稀少但评分却非常相似用户相似度较高的不合理现象，大大提高了邻居用户识别的准确性。

3.2 基于相关均值的推荐算法

在式(1)所示的传统推荐方法中，$\bar{U}_a$ 代表目标用户对已评分项目的平均评分，计算方法如下：

$$\bar{U}_a=\frac{1}{|I_a|}\sum_{i\in I_a}R_{a,i} \tag{4}$$

其中，$|I_a|$ 为目标用户 a 的已评分项目数；$R_{a,i}$ 表示用户 a 对项目 i ($i\in I_a$)的评分。这种均值计算方法的缺点在于过高估计了历史评分对预测的影响，因为历史均分高并不代表目标用户对当前项目就很感兴趣，历史均分低的情况下用户也可能对当前项目作出较高评价，所以不能有效反映该用户对未评分项目的喜好程度。本文对传统方法进行改进，充分利用目标用户和邻居间的相似性，以两者共同评价过的项目为基础计算的相关均值(Correlation Mean, CM)，代替单纯基于自身历史评分计算所得的均值。考虑到评分数据矩阵中目标用户和所有邻居 j 共同评分项目极少，而与其中某个邻居 j_i 共同评分项目又相对较多的实际情况，本文首先分别计算目标用户 a 和每个邻居 j_i 共同评价项目的评分均值，在此基础上进行加权平均，最终得到目标用户 a 的相关均值 $\bar{U}_{aCM}$，定义如下：

$$\bar{U}_{aCM}=\frac{1}{|NBS_a|}\sum_{j\in NBS_a}\left(\frac{1}{|I_{aj}|}\sum_{c\in I_{aj}}R_{a,c}\right) \tag{5}$$

其中，$|NBS_a|$ 表示用户 a 的邻居数目；I_{aj} 表示用户 a 和某个邻居 j($j\in NBS_a$)共同评分项目集；$R_{a,c}$ 为用户 a 对项目 c(c$\in I_{aj}$)的评分。该方法充分利用了目标用户和邻居间的相似性，以两者间共同评价项目为基础，所得均值更能准确反映目标用户对当前项目的真实喜好，尽管此方法需要多次计算均值，但仅针对较少的共同评分项目而非整个项目空间，因此计算复杂度相比传统方法略有提高。

在本文提出的相似性度量以及相关均值计算方法基础上，按如下公式预测目标用户 a 对未评分项目 i 的评分 $P_{a,i}$ ：

$$P_{a,i}=\bar{U}_{aCM}+\frac{\sum_{j\in NBS_a}sim'(a,j)\times(R_{j,i}-\bar{R}_j)}{\sum_{j\in NBS_a}|sim'(a,j)|} \tag{6}$$

其中，$\bar{U}_{aCM}$ 表示目标用户 a 的相关均值；$sim'(a,j)$ 表示利用加权 Pearson 相关系数度量所得的用户 a 和用户 j 的相似度；$R_{j,i}$ 为邻居 j 对项目 i 的评分；$\bar{R}_j$ 为邻居 j 对项目的平均评分。整个推荐算法的具体描述如下：

算法 1 基于相关均值的协同过滤算法 $CMA(a,\boldsymbol{R},k,N)$

输入 目标用户 a ；用户-项目评分矩阵 $\boldsymbol{R}$；邻居数目 k ；推荐项目数 N 。

输出 目标用户的 $top-N$ 推荐集。

Step1 从评分矩阵 $\boldsymbol{R}$ 中检索出所有 m 个用户、 n 个项目及目标用户 a 的已评分项目集，分别记为 $U_m=\{u_1,u_2,\cdots,u_m\}$，$I_n=\{i_1,i_2,\cdots,i_n\}$， $I_a=\{i_1,i_2,\cdots,i_p\}$ 。

Setp2 对任意用户 $u\in U_m(u\neq a)$ ，读取 $\boldsymbol{R}$ 中数据，根据式(2)计算它与目标用户 a 的相似度 $sim'(a,u_i)$ ，并选择相似度较大的 k 个用户作为 a 的最近邻集 $NBS_a=\{j_1,j_2,\cdots,j_k\}$ 。

Step3 根据式(5)计算目标用户 a 相对其邻居的相关均值 $\bar{U}_{aCM}$ 。

Step4 对每个项目 $i\in I_c$ $(I_c=I_n-I_a)$ ，根据式(6)计算 i 对 a 的推荐度 $P_{a,i}$ 。

Step5 将 I_c 中的项目按推荐度大小排列，选择其中最前的 N 个项目作为目标用户 a 的 $top-N$ 推荐集。

4 实验结果及其分析

4.1 数据集

本文采用 MovieLens 站点提供的数据作为数据集，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推荐列表。目前该站点用户已超过 7 万人，用户已评分电影超过 5 000 部。本文对其中的用户评分数据库进行筛选，选择了 10 000 条评分数据作为实验数据集，其中包含 212 个用户和 986 部电影，且每个用户至少对 20 部以上的电影进行了评分。并将数据集进一步划分为 80%的训练集和 20%的测试集。

4.2 评价标准

算法精确度是评价推荐算法性能的一个重要指标，通常采用易于理解的平均绝对误差(Mean Absolute Error, MAE)作为衡量推荐系统质量的一个度量标准[4]。通过计算算法所预测的用户评分与实际的用户评分间的偏差来度量预测的准确性，平均绝对误差越小，推荐质量越高。

设经算法预测的 $top-N$ 推荐集的评分为 $p_1, p_2, \cdots, p_N$，相应的实际评分为 $q_1, q_2, \cdots, q_N$，则 MAE 定义如下：

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{7}$$

4.3 实验结果及分析

4.3.1 相似性度量标准比较

由于相似性度量是邻居形成过程中的关键步骤，直接影响到推荐算法的准确性，因此本实验首先在同一推荐算法前提下，比较传统相似性度量方法及本文加权 Person 相关系数方法对推荐质量的影响。以上述数据集为基础，分别计算 3 种方法下基于相关均值推荐算法的平均绝对误差。实验结果如图 1 所示。

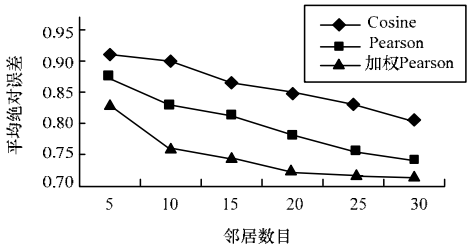

图 1　3 种相似性度量方法的平均绝对误差

由图 1 可知，在不同邻居数目下，加权 Person 相关系数度量方法的 MAE 值均低于其他 2 种传统度量方法，可见基于用户共同评分项目数的权重因子的引入，有效提高了用户相似性度量的准确性。

为简化计算，上述实验中加权 Person 相关系数方法中的估计均值 D 一旦设定就不再变化，而不同电子商务系统中数据的稀疏状况存在较大差异。图 2 给出了 D 的不同取值(5, 10, 20, 50)对推荐精度的影响。

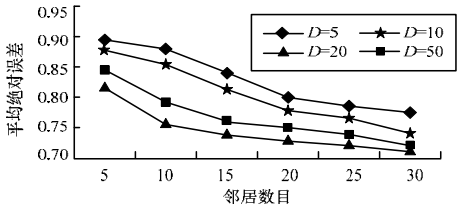

图 2　加权 Person 方法中 D 的取值对算法性能的影响

从中可以发现，当设定的 D 值过小(D=5)，数据集中大部分用户两两间共同评分的项目数往往都会大于 D 值，根据式(3)计算所得的权重都为 1，这样就等效于传统 Person 方法，无法提高邻居用户的识别准确性。因此，可以针对实际情况动态调整 D 值来优化推荐效果，针对本文数据集，D 值取 20 时可获得较好的推荐质量。

4.3.2 推荐算法性能比较

为检验本文提出的基于相关均值的推荐算法(CMA)的推荐效果，将该算法与传统的基于最近邻协同过滤算法(NBCFA)在相同数据集上进行对比实验，结果如图 3 所示。

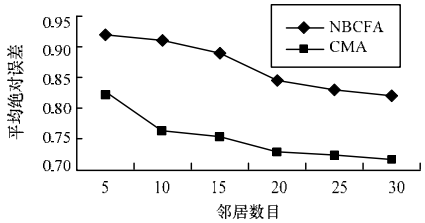

图 3　2 种推荐算法精确度比较

由图 3 可知，针对相似性度量和目标用户均值计算方法作了改进的本文算法在不同邻居数目下的精确度均优于传统方法。

5 结束语

本文分析了传统基于最近邻协同过滤推荐算法在评分数据稀疏情况下存在的问题，采用加权 Person 相关系数度量用户间的相似性，使邻居用户的识别更加准确有效，并基于邻居共同评分项目计算目标用户的相关均值。实验结果表明，在此基础上改进的推荐算法显著提高了系统的推荐质量。

参考文献

[1] Goldberg D, Nichols D. Using Collaborative Filtering to Weave an Information Tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[2] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003, 41(9): 1621-1628.

[3] 周军锋, 汤　显, 郭景锋. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 2004, 41(10): 1842-1847.

[4] Sarwar B, Karypis G, Konstan J, et al. Item-based Collaborative Filtering Recommendation Algorithms[C]//Proc. of the 10th International World Wide Web Conference. Hong Kong, China: [s. n.], 2001.

编辑　金胡考

(上接第 52 页)

参考文献

[1] Motwani R, Widom J, Arasu A, et al. Query Processing, Approximation, and Resource Management in a Data Stream Management System[C]//Proc. of the 1st Biennial Conf. on Innovative Data Systems Research. Asilomar, USA: [s. n.], 2003.

[2] Tatbul N, Cetintemel U, Zdonik S, et al. Load Shedding in a Data Stream Manager[C]//Proc. of the 29th Int'l Conf. on Very Large Data Bases. Berlin, Germany: [s. n.], 2003.

[3] Babcock B, Datar M, Motwani R. Load Shedding for Aggregation Queries over Data Streams[C]//Proc. of the 20th International Conference on Data Engineering. Boston, USA: [s. n.], 2004.

[4] Kang J, Naughton J, Viglas S. Evaluating Window Joins over Unbounded Streams[C]//Proc. of the 19th Int'l Conf. on Data Engineering. [S. l.]: ACM Press, 2003.

[5] 王伟平, 李建中, 张冬冬, 等. 基于滑动窗口的数据流连续 J-A 查询的处理方法[J]. 软件学报, 2006, 17(4): 740-749.

[6] 刘学军, 胡　平, 徐宏炳, 等. 基于滑动窗口的在线数据流增量聚集查询[J]. 计算机工程, 2007, 33(21): 45-49.

编辑　张　帆