第37卷 第19期 计 算 机 工 程 2011年10月
Vol.37 No.19 Computer Engineering October 2011

·软件技术与数据库· 文章编号：1000—3428(2011)19—0059—03 文献标识码：A 中图分类号：TP311

基于单分类的协同过滤推荐算法

杨 帅，薛 文，谢永红，王晓宇，祝小杰

(北京科技大学信息工程学院，北京 100083)

摘 要：随着电子商务推荐系统中用户和商品数目的增加，用户商品评分数据集的稀疏性会导致协同过滤推荐算法的推荐质量下降。针对该问题，提出一种基于单分类的协同过滤推荐算法。根据目标用户评分商品对应的类别，选择候选最近邻居集，采用单分类预测用户对商品的评分，以减小目标用户与候选最近邻居所形成的数据集稀疏性。实验结果表明，该算法能提高寻找最近邻居的准确性，从而改善协同过滤的推荐质量。

关键词：推荐系统；协同过滤；数据稀疏性；单分类；平均绝对偏差

Collaborative Filtering Recommendation Algorithm Based on Single-class Classification

YANG Shuai, XUE Wen, XIE Yong-hong, WANG Xiao-yu, ZHU Xiao-jie

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

【Abstract】With the increasing number of users and goods in E-commerce recommender systems, the data set sparse of user goods rating reduces the quality recommendation of collaborative filtering recommendation algorithm. To solve this problem, this paper proposes a collaborateive filtering recommendation algorithms based on single-class classificatin. It chooses candidate nearest neighbor set which depending on the target user rating goods corresponding to category and uses single-class classification to predict the values of the user rating. It can reduce the sparse of data set which is formed by the target user and the candidate nearest. Experimental results show that the algorithm is able to increase the accuracy of searching nearest neighbor set, resulting in improving recommendation quality of the collaborative filtering.

【Key words】recommendation system; collaborative filtering; data sparse; single-class classification; Mean Absolute Error(MAE)

DOI: 10.3969/j.issn.1000-3428.2011.19.018

1 概述

当今，随着互联网的普及和电子商务的蓬勃发展，个性化推荐作为电子商务的重要组成部分，已成为研究的热点。协同过滤作为目前最广泛使用的推荐算法，其基本思想是：根据目标用户的历史评分，寻找目标用户的最近邻居用户，根据最近邻居对商品的评分计算目标用户对商品的评分值，选择评分最高的前 N 项商品集合作为推荐集反馈给目标用户[1]。因此，用户对商品评分的数据收集越多，协同过滤算法的推荐质量越高。但是，随着电子商务站点用户和商品数量的不断增加，用户商品评分数据集的稀疏性导致协同过滤算法的推荐质量下降。对此，已有一些改进方法，常见的简单方法是将一个固定的缺省值填充所有未评分项[2]，但用户对未评分项的评分不可能完全相同；还有采用奇异值分解的降维技术[3]，通过减少用户商品评分数据集的维数达到降低数据集稀疏性，但这种降维方法会导致信息丢失，对推荐的质量产生负面影响。

本文在传统协同过滤算法的基础上，使用单分类[4]对未评分项进行预测，填充未评分项，平滑数据集，解决了数据集稀疏的问题，使得交叉评分增加，从而提高寻找最近邻居的准确度。

2 相关工作

2.1 用户商品评分矩阵

用户商品评分矩阵 $\boldsymbol{R}_{m\times n}$ 是一个 m 行 n 列的矩阵，用户 $u_i \in U=\{u_1,u_2,\cdots,u_m\}$，$m$ 为用户的个数，U 为用户集合；商品 $i_j \in I=\{i_1,i_2,\cdots,i_n\}$，$n$ 为商品的个数，I 为商品集合；$R_{i,j} \in \boldsymbol{R}_{m\times n}$，$R_{i,j}$ 为用户 i 对商品 j 的评价值。

2.2 商品类别

在电子商务网站中，所有商品项可以划分到有限的若干个商品类别中，例如，当当网将图书分为文学、计算机、管理等多种类别。由此，商品集合 I 中的商品项 i_j 是属于类别集合 C 中的 C_j。商品的所属类别可通过商品属性表获得，商品属性表包括商品的基本信息如价格、品牌、类别等属性，商品各属性的获得可以由专家给出，也可以通过数据库和Web日志生成。

2.3 相似性度量

为了找到目标用户的最近邻居，必须度量用户之间的相似性，然后选择相似性最高的若干用户作为目标用户的最近邻居。度量用户之间相似性的方法[5]主要有：

(1)余弦相似性。用户评分被看作是 n 维商品空间上的向量，用户间的相似性通过向量间的余弦夹角度量。设用户 u 和用户 v 在 n 维商品空间上的评分分别表示为向量 $\boldsymbol{u}'$、$\boldsymbol{v}'$，则用户 u 和用户 v 之间的相似性计算方法如下：

基金项目：国家自然科学基金资助项目(60675030, 60875029)

作者简介：杨 帅(1986－)，男，硕士研究生，主研方向：数据挖掘，个性化推荐；薛 文，硕士研究生；谢永红，副教授；王晓宇、祝小杰，硕士研究生

收稿日期：2011-03-24 **E-mail：**yangshuai720@gmail.com

$$sim(u,v)=\cos(u,v)=\frac{\boldsymbol{u}'\cdot\boldsymbol{v}'}{|\boldsymbol{u}'|\times|\boldsymbol{v}'|} \tag{1}$$

(2)修正的余弦相似性。在余弦相似性度量方法中没有考虑不同用户的评分尺度问题，修正的余弦相似性度量方法通过减去用户对商品平均评分来改善上述缺陷。设用户 u 和用户 v 共同评分的商品集合 $I_{u,v}$，I_u 和 I_v 分别表示用户 u 和用户 v 评分的商品集合，则用户 u 和用户 v 之间的相似性计算方法如下：

$$sim(u,v)=\frac{\sum_{k\in I_{u,v}}(R_{u,k}-\overline{R}_u)(R_{v,k}-\overline{R}_v)}{\sqrt{\sum_{k\in I_u}(R_{u,k}-\overline{R}_u)^2}\sqrt{\sum_{k\in I_v}(R_{v,k}-\overline{R}_v)^2}} \tag{2}$$

其中，$R_{u,k}$ 和 $R_{v,k}$ 分别表示用户 u、v 对商品 k 的评分；$\overline{R}_u$ 和 $\overline{R}_v$ 分别表示用户 u 和用户 v 对商品的平均评分。

2.4 预测计算

通过相似性计算得到用户 u 的最近邻居集合，表示为 NBS_u。根据 NBS_u，用户 u 对商品 i 的预测评分，可通过如下公式计算：

$$P_{u,i}=\overline{R}_u+\frac{\sum_{v\in NBS_u}sim(u,v)\times(R_{v,i}-\overline{R}_v)}{\sum_{v\in NBS_u}sim(u,v)} \tag{3}$$

3 算法设计

通过对传统协同过滤推荐算法研究分析，发现存在的问题主要为：

(1)目标用户最近邻居的搜索是在整个用户空间上，即候选最近邻居集是整个用户空间；

(2)在实际应用中，每个用户对商品的评价信息量是有限的，使得目标用户与候选最近邻居集形成的评分矩阵稀疏。

可通过减少目标用户的候选最近邻居数目来解决以上问题，同时，采用单分类方法对用户的未评分项进行预测评分，添加到目标用户与候选最近邻居集形成的评分矩阵中。经过填充的矩阵，数据的稀疏性得到改善，用协同过滤算法对填充后的矩阵进行运算，提高了寻找最近邻居的准确度，有效地解决在数据稀疏情况下传统的协同过滤算法存在的不足，从而提高推荐质量。

3.1 候选最近邻居集

相关数据表明，一般用户最多不过就购买了整个网站商品的 1%~2%[5]，但根据其兴趣偏好用户购买的商品基本集中在某几个类别中，因此，商品类别实质上对应着用户的兴趣领域。用户可能由于搜索、获知途径等因素的不同，分别对同一类别下的不同的商品项进行了评分，但用户在这一兴趣领域的偏好是相同，因此，找出在相同类别上具有商品项评分的用户组成候选最近邻居集是合理的。

定义(候选最近邻居集) 设目标用户 u 的评分商品对应的类别集合为 $C_t=\{c_1,c_2,\cdots,c_g\}$，则对于 $\forall c_p\in C_t\ (1\leqslant p\leqslant g)$，择取用户商品评分矩阵 $\boldsymbol{R}_{m\times n}$ 中所有对于 c_p 相应评分的用户集 U_p 以及属于 c_p 的商品集 I_p，组成 u 的候选最近邻居集为 $CNBS_u=U_p-u$，同时得到 u 与候选最近邻居集形成的评分矩阵 $\boldsymbol{R}_p=U_p\times I_p$。

3.2 单分类预测未评分项

对于目标用户 u 与候选最近邻居集形成的评分矩阵 $\boldsymbol{R}_p$ 中未评分项的填充可通过分类学习进行预测，对于分类学习，需要正例和反例，即用户感兴趣和不感兴趣 2 类样本建立分类器。在实际应用中，大多数电子商务网站是在用户购买完商品后才能进行评分，一般用户感兴趣才会去购买，购买完后再去评分就使得评分值的高低对表现用户是否感兴趣意义不是很大了。正例的获取通过用户评分过的商品，而反例的获取既不能通过评分值的高低来判断也不能简单的认为用户没有评分就不感兴趣，确实是可以要求用户在浏览网页商品的过程中对商品去标识是否感兴趣。但这种方法会妨碍用户的正常浏览遭到用户反感。对于用户不感兴趣的商品获取难，因此，只能通过用户的一些已知兴趣这一类样本来进行单分类学习建立分类器，对 $\boldsymbol{R}_P$ 中的未评分项进行预测，算法如下：

输入 目标用户与候选最近邻居集的评分矩阵 $\boldsymbol{R}_p=U_p\times I_p$，商品属性 T

输出 用户预测评分集合 Q

```
1  i=1, search u_i in 用户集 U_p ;
2  while( u_i ≠ Null){
3    在 R_p 中，Search u_i 的已评分商品集合
       I_{u_i} ∈ I_p;
4    形成用户 u_i 的正例集合 Z= I_{u_i} ∩ T;
5    在 R_p 中，Search u_i 的未评分商品集合
       L_{u_i} ∈ I_p，其中 L_{u_i} =I_p–I_{u_i};
6    形成用户 u_i 的未标识集合 W= L_{u_i} ∩ T;
7    根据 Z 与 W 采用单分类方法建立分类器;
8    对 W 进行分类预测；}
9  i++;
10 return 用户预测评分集合 Q。
```

对于步骤 7 解释如下：采用的单分类学习方法为合成两阶段法，该方法不需要反例的获取，只根据正例和未标识样例集合建立分类器，不仅减少了对训练集中标识类别数据的要求，还利用大量廉价易得的未标识数据辅助分类。

第 1 阶段根据正例和未标识样例集合获取可靠反例，采用的方法是未标识集合中的任意样例 w_d 在 Z 中找到与此最近的一个正例 z_d，计算之间的距离。根据 k 值，在 Z 中找到与 z_d 最近的 k 个正例，计算出 z_d 与它们之间的距离，如果两个距离相比小于某一设定阈值 δ，则判断 w_d 属于反例，对于其中的距离计算是各属性之间的欧式距离。第 1 阶段算法如下：

输入 Z, W

输出 可靠反例 KF

```
1  设置一个值域δ和 k 的值;
2  为了识别 ∀ w_d∈W 是否为反例，首先找到与 w_d 最近正例
     z_d∈Z，计算出 w_d 与 z_d 之间的距离 D_1;
3  If  k=1
     计算 z_d 和其最近的一个正例之间的距离 D_2;
   else
     计算 z_d 与其最近 k 个正例之间的距离 D_2;
4  If  D_1 / D_2>δ
     w_d∈KF;
   else
     w_d ∉ KF;
```

第 2 阶段根据正例和获取的可靠反例集合建立初始的 SVM 分类器，对 W 进行分类，分为反例的添加到可靠反例中，迭代建立分类器，在每次迭代中，对未标记样例集合循序渐渐的学习和标记，随着未标记集合规模的减小，标记样本数量的增大，直到未标记集合中无样本可标记，最终完成对未标记集合的分类，形成最终分类器。第 2 阶段算法如下：

输入 Z, W, KF

输出 最终分类器 SVM_{last}

1 使用 Z 和 KF 训练新的 SVM 分类器 SVM_i；
2 用分类器 SVM_i 对 Q(W-KF)进行分类；
3 对 Q 中分类为反例的样本，组成集合 M；
4 Q = Q-M；KF=KF∪M；
5 goto(3) until W ==null；
6 return 最终分类器 SVM_{last}。

3.3 矩阵填充

对目标用户 u 与候选最近邻居集形成的评分矩阵 $\boldsymbol{R}_p$ 中的未评分项进行预测后，为填充评分矩阵提供了依据。经过填充的矩阵，使其数据的稀疏性得到了改善。算法如下：

输入 目标用户与候选最近邻居集的评分矩阵 $\boldsymbol{R}_p=U_p\times I_p$，用户的预测评分集合 $\boldsymbol{Q}$

输出 填充矩阵 $\boldsymbol{R}'_P$

1 i=1，search u_i in 用户集 U_p；
2 While($u_i\neq$Null){
3 get u_i 对应的预测评分集合 $Q_{u_i}\in Q$；
4 While($q_j\in Q_{u_i}$ && $q_j\neq$Null){
5 get u_i 对 q_j 的 R_{ij}；
6 Fill matrix R_p；
7 j++；}
8 i++；}
9 return 填充矩阵 R'_p。

3.4 推荐产生

完成对评价矩阵的填充后，用户间的评分商品增多，按照修正余弦法(式(2))求出目标用户与候选最近邻居集的相似性得到最近邻居集，通过式(3)计算目标用户对商品的预测评分值，按值从大到小取前 N 个组成 TOP-N[1]推荐集 I_{rec}，推荐给目标用户，从而完成整个推荐过程。

输入 目标用户 u，目标用户与候选最近邻居集的未填充评分矩阵 $\boldsymbol{R}_p=U_p\times I_p$，填充后矩阵 $\boldsymbol{R}'_p$

输出 目标用户 u 的推荐集 I_{rec}

1 i=1，search u_i in 用户集 U_p；
2 while ($u_i\neq$Null && $u_i\neq u$){
3 Search $I'_{u,u_i}=I'_u\cap I'_{u_i}$ in R'_p；//在填充后的矩阵 R'_p 中查找
//用户 u 和 u_i 共同评价的商品集合
4 $sim(u,u_i)$=adjusted-cosine(u,u_i) in I'_{u,u_i}；//根据式(2)计算用
//户 u 和 u_i 之间的相似度
5 按数值大小，将 $sim(u,u_i)$添加到用户相似性数组 U_{sim}；
6 i++；}
7 选择 Top K in U_{sim}，构成最近邻居集 NBS_u；
8 在 R_p 中，Search u 的未评分商品集合 L_u；
9 j=1，search $(L_u)_j$ in 未评分商品集合 L_u；
10 while ($(L_u)_j\neq$Null) {
11 $R_{uj}=P_{u,j}$；//根据式(3)计算用户 u 对商品 j 的预测评分
12 j++；}
13 选择 Top N in R_{uj}；
14 return 目标用户 u 的推荐集 I_{rec}。

4 实验及分析

4.1 实验环境与数据集

实验硬件环境为联想 2.97 GHz 个人计算机，实验的主要软件环境为 Microsoft Windows XP、SQL Server 2005、VisualC++6。实验数据集取自 MovieLen 站点，包括 943 位用户对 1 682 部电影的 10 万条评分数据，每位用户至少对 20 部电影进行了评分，所有电影分属于 19 种电影类别，同时，还整理出这 1 682 部电影的 20 个特征属性作为其属性表。从该数据集中随机选择 80 位用户对 1 119 部电影的 7 632 条评分数据作为实验数据集。实验分别随机抽取每位用户的 6 条评分数据组成测试集，余下的 7 143 条评分数据作为训练集。为度量整个数据集的稀疏性，引入稀疏等级[5]的概念，其定义为用户评分数据矩阵中未评分条目所占的百分比，选择的电影数据集的稀疏等级为：1−7 632/(80×1 119)=0.914 7。

4.2 度量标准

评价推荐系统推荐质量的度量标准主要有统计精度度量方法和决策支持精度度量方法[5]。实验采用统计精度度量方法中广泛使用的平均绝对偏差(Mean Absolute Error, MAE)。MAE 越小评分预测越准确，推荐质量越高。设推荐集中的项目评分预测值为$\{p_1, p_2,\cdots,p_N\}$，目标用户对这些项目的实际评分为$\{q_1,q_2,\cdots,q_N\}$，则 MAE 为：

$$M_{\mathrm{MAE}}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N},\ i=1,2,\cdots,N$$

4.3 结果分析

为了检验本文提出算法的有效性，把本文算法和一般的基于余弦、基于修正余弦的协同过滤算法作比较，计算在不同最近邻居时各种推荐算法的 MAE，实验结果如图 1 所示。

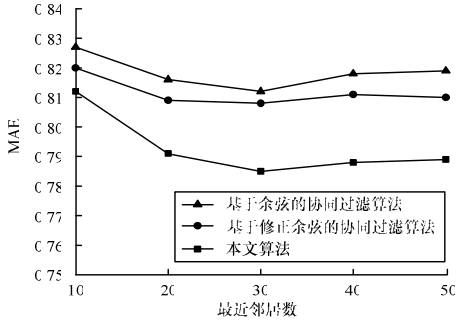

图 1 MAE 随最近邻居数的变化

可以看出，在最近邻居个数不同情况下，本文算法具有最小的 MAE 值。因此，在用户评分数据极端稀疏的情况下，本文算法具有更好的性能表现，证明单分类能够缓解协同过滤的稀疏性。

5 结束语

本文分析了用户商品评分数据稀疏情况下协同过滤算法存在的问题，提出一种基于单分类的协同过滤算法。实验结果表明该算法能提高系统的推荐质量。虽然本文算法延长了找到最近邻居的时间，但这个过程可以离线计算。未来的工作将在推荐质量和推荐系统的实时性上寻找一个平衡点。

参考文献

[1] Karypis G. Evaluation of Item-based Top-n Recommendation Algorithms[C]//Proc. of the 10th International Conf. on Information and Knowledge Management. New York, USA: ACM Press, 2001.

[2] Dasu T, Johnson T. Exploratory Data Mining and Data Cleaning[M]. [S. l.]: Wiley Press, 2003.

[3] Sarwar B M, Karypis G, Konstan J A, et a1. Application of Dimensionality Reduction in Recommender System: A Case Study[R]. Minneapolis, USA: Department of Computer Science and Engineering, University of Minnesota, Tech. Rep.: TR 00-043, 2000.

[4] 陈志敏，沈 洁，赵 耀. 基于相关均值的协同过滤推荐算法[J]. 计算机工程, 2009, 35(22): 53-55.

[5] 黄国言，李有超. 基于多序选择域的协同过滤推荐算法[J]. 计算机工程, 2010, 36(7): 36-38.

编辑 顾姣健