# 二分网络中基于谱聚类的协同推荐

张思明，游天童
（福州大学 数学与计算机学院，福建 福州 350108）

**摘 要**：提出一种基于谱聚类的协同推荐算法(SCBCF)。首先从用户——项目二分网络的单顶点投影中得到用户之间的相似矩阵，然后对该矩阵应用谱聚类算法，将用户聚成 $k$ 类，并将得到的聚类结果用于数据平滑和邻居结点的选择，最后基于最近邻居集评分行为，对目标用户产生推荐。在 MovieLens 上的实验结果证明本文方法比传统的协同过滤算法能更好地应用于二分网络的协同推荐。

**关键词**：协同过滤；谱聚类；推荐算法；平均绝对偏差
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1674-7720(2012)22-0060-04

## Collaborative recommendation based on spectral clustering in bipartite network

Zhang Siming, You Tiantong
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

**Abstract**: This paper proposes an approach based on spectral clustering for collaborative recommendation. Firstly, we conduct user-user similarity matrix from item-user bipartite network through one-mode projection. Then we apply spectral clustering method to cluster users into $k$ clusters from the user-user similarity matrix for data smoothing and neighborhood selection. At last, we make recommendation for target user based on the rating behavior of nearest neighbors set. Experiment result on MovieLens shows that our proposed approach is better than traditional collaborative filtering algorithm.

**Key words**: collaborative filtering; spectral clustering; recommendation algorithm; MAE

随着互联网信息的不断膨胀，信息过载也越来越严重，因而推荐系统越来越受到人们的重视。最简单的推荐算法是全局排名方法 GRM(Global Ranking Method)，该算法不考虑用户的个性化需求，因而其推荐结果的质量并不好。于是，考虑用户偏好的协同过滤 CF(Collaborative Filtering)推荐算法被广为应用，并迅速成为最受欢迎的推荐算法之一。协同过滤算法考虑用户兴趣，在用户群中寻找目标用户的相似用户组，综合这些相似用户对某一项目的评价，预测目标用户对此项目的兴趣。

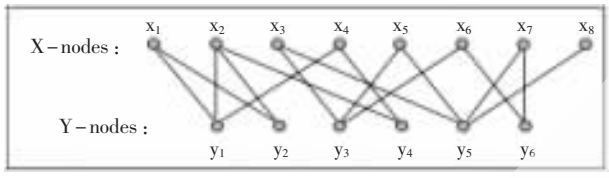
目前，协同过滤算法主要分为两类[1]：基于内存的方法和基于模型的方法。基于内存的方法在整个数据库上执行，从训练数据库中找出与目标用户最相关的 $K$ 个用户，然后把他们的评分信息结合在一起对目标用户的评分情况进行预测。主要有基于 Pearson 相关性的方法、基于向量相似度的方法等。这些算法主要有两个缺点：易受稀疏的评分数据的影响；算法的可伸缩性差。与之相对，基于模型的方法并不直接使用单个用户的评分信息，而是预先按照用户评分的模式对用户进行聚类，然后计算目标用户与各个类别之间的相似度，找出最相似的类，用该类对某个项目的评分作为目标用户对该项目的评分。主要的方法有贝叶斯网络方法、聚类的方法。基于模型的方法在建立聚类的过程中较为耗时，而且对目标用户做出的评分预测也存在准确性较低的问题。

本文考虑将谱聚类的方法引入到协同过滤推荐算法中，对训练集中的用户进行谱聚类，结合基于内存和基于模型这两种方法的优势，而对目标用户评分的预测任务则交由其最相关的用户群组来完成。对于如何构造谱聚类算法的输入矩阵，本文将用户——项目二分网络投影到只包含用户结点的单顶点网络，构造 $n\times n$ 的用户相似矩阵。考虑到评分数据的稀疏性，本文利用类的信息对类中每个用户未评分的项目进行数据平滑处理，从得到的 $N$ 个聚类中找出与目标用户最相似的一个或几个类别作为最近邻居候选集，再从候选集中找出最相似的 $K$ 个用户进入最近邻居集，最后预测目标用户对每个项目的评分。

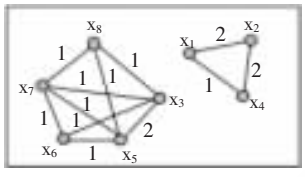
## 1 相关工作

### 1.1 二分网络的投影

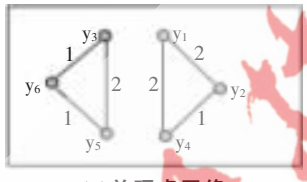
二分网络单顶点投影[2]是研究二分网络的一个重要方法。二分网络投影成单顶点网络的方式主要分为两类：无权投影和加权投影。如图1所示，图1(b)、图1(c)分别为图1(a)关于$X$、$Y$的单顶点投影，单顶点网络中的任意两个点之间边的权值大小为这两点在二分网络中的共同邻居数。虽然单顶点网络无法完全描述二分网络的全部信息，但是这个只含一种结点的单结点网络完美地保存了二分网络中此类结点的拓扑关系，网络中边的权值构成的关系矩阵可以用来表示同类结点之间的相似关系。




(a)二分网络




(b)单顶点网络




(c)单顶点网络

图1 二分网络及其$X$投影和$Y$投影图

### 1.2 谱聚类

近年来，谱聚类已经成为一种广受欢迎的现代聚类算法[3]。与传统的聚类算法如K-means算法相比，这种算法效率更高，聚类结果更优。谱聚类易于实现，可以用标准的线性代数的方法来高效解决。

给定数据点集$x_1, x_2, \cdots, x_n$，以及所有点对$x_i$和$x_j$之间的相似度$s_{ij} \geqslant 0$所构成的$n \times n$的相似度矩阵$S$。聚类的目标是把这些点划分进不同的类中，使得类内点的相似度大，而类间点的相似度小。本文用一个相似图$G=(V,E)$来表示上述信息，每个顶点$v_i$表示数据点$x_i$。如果$s_{ij} \geqslant 0$，那么顶点$v_i$和$v_j$之间存在边，且其边的权值即为$w_{ij}$。将二分网络抽象成二分图之后，可以这样来阐述聚类问题：找到一个图的划分，使得不同组之间的边的权值和小，而组内边的权值和大。

谱聚类的衍化算法有很多种，此处介绍的是非正规化谱聚类算法，这也是本文在后面推荐算法中用到的。非正规化的谱聚类算法描述如下：

输入：相似度矩阵$\boldsymbol{S} \in \boldsymbol{R}^{n \times n}$，聚类数$k$；

输出：$k$个聚类$A_1, A_2, \cdots A_k$，$A_i=\{j \mid y_j \in C_i\}$；

function spectralClustering($\boldsymbol{W}, k$)

$\boldsymbol{D}=\mathrm{diag}(\mathrm{sum}(\boldsymbol{W}))$；//$\boldsymbol{D}$为对角矩阵，$d_i=\sum_j w_{ij}$

$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$；//$\boldsymbol{L}$为非正规化的Laplacian矩阵

$\begin{bmatrix}\boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_k\end{bmatrix}=\mathrm{eigenvectors}(\boldsymbol{L}, k)$；//计算$\boldsymbol{L}$的前$k$个最小特征值对应的特征向量$\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_k$

$\boldsymbol{U}=(\boldsymbol{u}_1^{-1}, \cdots, \boldsymbol{u}_k^{-1})$；//以$\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_k$为列向量组成矩阵$\boldsymbol{U} \in \boldsymbol{R}^{n \times k}$

$\mathrm{kmeans}(\boldsymbol{U}, k)$；//用K-means算法将$n \times k$维的数据点$y_i \in R^k$聚到$k$个类中$C_1, \cdots, C_k$

end

其中算法的输入矩阵$S$，可以用二分图的邻接矩阵$W$来近似表示。

### 1.3 协同过滤

作为构建推荐系统的最成功方法之一，协同过滤使用已知用户群组的偏好来预测与该群组相似的其他用户的未知偏好。协同过滤基于以下两个假设：(1)人与人之间存在偏好兴趣上的某种程度的相似或者重叠；(2)人对事物的偏好是具有稳定性的。因此，寻找目标用户的最相似的用户群组对协同过滤推荐来说是至关重要的。一般来说，协同过滤的步骤为：(1)构建用户档案，即收集用户的评分、评价行为等，并进行数据清理、转换，最终形成用户——项目的评价矩阵；(2)最近邻居搜索，即计算目标用户与数据库内各个用户的相似度，找出相似度最高的$N$个用户作为最近邻居集；(3)推荐产生，即根据最近邻居集的评价值产生推荐。

用户相似度的主要度量方式有Pearson相关系数法、余弦相关性法和修正的余弦相关性法[4]。

Pearson相关系数：设经用户$i$和用户$j$共同评分的项目集合用$I_{ij}$表示，则用户$i$和用户$j$之间的相似性$\mathrm{sim}(i,j)$通过Pearson相关系数度量为：

$$\mathrm{sim}(i,j)=\frac{\sum_{c \in I_{ij}}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c \in I_{ij}}(R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c \in I_{ij}}(R_{j,c}-\overline{R_j})^2}} \tag{1}$$

其中，$R_{i,c}$表示用户$i$对项目$c$的评分，$\overline{R_i}$和$\overline{R_j}$分别表示用户$i$和用户$j$对项目的平均评分。

余弦相似性(cosine)：把用户评分看做是$n$维项目空间上的向量，用户间的相似性通过向量间的余弦夹角来度量。设用户$i$和用户$j$在$n$维项目空间上的评分分别为向量$\vec{i}$和$\vec{j}$，则用户$i$和用户$j$之间的相似性$\mathrm{sim}(i,j)$为：

$$\mathrm{sim}(i,j)=\cos(\vec{i},\vec{j})=\frac{\vec{i}\cdot\vec{j}}{|\vec{i}||\vec{j}|} \tag{2}$$

其中，分子为两个用户评分向量的内积，分母为两个用户向量模的乘积。

修正的余弦相似性(adjusted cosine)：修正的余弦相似性度量方法考虑不同用户的评分尺度问题。设经用户$i$和用户$j$共同评分的项目集合用$I_{ij}$表示，$I_i$和$I_j$分别表示经用户$i$和用户$j$评分的项目集合，则用户$i$和用户$j$

之间的相似性 $\mathrm{sim}(i,j)$ 为：

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c\in I_j}(R_{j,c}-\overline{R_j})^2}} \tag{3}$$

其中，$R_{i,c}$ 表示用户 $i$ 对项目 $c$ 的评分，$\overline{R_i}$ 和 $\overline{R_j}$ 分别表示用户 $i$ 和用户 $j$ 对项目的平均评分。

## 2 SCBCF 算法

本文将基于谱聚类的协同过滤算法记为 SCBCF 算法，以方便说明。

### 2.1 问题定义

给定一个二分图 $G=\langle T\cup U,E\rangle$，其中 $T=\{t_i|1\leqslant i\leqslant m\}$，$U=\{u_i|1\leqslant i\leqslant n\}$。这里用 $T$ 表示项目集，$U$ 表示用户集。同类结点之间不允许存在边。二分图 $G$ 可以用 $m\times n$ 的矩阵 $W$ 来表示，$W(i,j)$ 表示边 $\langle i,j\rangle$ 的权值。现在的任务是，给定一个目标用户 $u_a$，预测其在 $T$ 上已经评价过的项目的评分，并与 $u_a$ 对该项目的实际评分相比较，以此来验证算法的准确性。

### 2.2 SCBCF 算法框架

预处理阶段：用 1.2 节中给出的算法对用户集进行聚类，形成 $k$ 个类别：$C_1,\cdots,C_k$。

推荐阶段：给定一个目标用户 $u_a$，项目 $t$，最近邻居数 $K$。

(1)从谱聚类的结果中，选择几个与 $u_a$ 最相似的类，作为最近邻居的候选集。

(2)对候选集中的每个用户，计算其与 $u_a$ 的相似度 $\mathrm{sim}(u_a,u)$，其中 $u$ 的评分结合了 $R_u(t)$ 和 $R_{C_u}(t)$，前者表示 $u$ 对 $t$ 的评分，后者表示 $u$ 所在类对 $t$ 的类评分。

(3)选择 $K$ 个最相似的用户作为最近邻居。

(4)基于最近邻居的评价行为，预测 $u_a$ 对项目 $t$ 的评分。

#### 2.2.1 聚类算法

本文采用 1.2 节中的非正规化的谱聚类算法对训练集中的用户进行聚类。聚类之前，根据图 1 中的形式，先将二分图投影成只含用户的单顶点网络，顶点之间的权值即为这两个用户收集的共同项目数，所有顶点之间的权值构成权值矩阵 $W$，作为聚类算法中的输入矩阵。

#### 2.2.2 数据平滑处理

正如前面所提到的，数据的稀疏性是协同过滤面临的一个基本问题。本文应用基于聚类的数据平滑策略[5]。以下定义一个特别的评分值：

$$R_u(t)=\begin{cases}R_u(t) & t\in T_u\\ \hat{R}_u(t) & t\notin T_u\end{cases} \tag{4}$$

其中，$T_u$ 表示 $u$ 收集的项目集，$\hat{R}_u(t)$ 表示 $u$ 对未收集的项目 $t$ 的平滑评分值。$\hat{R}_u(t)$ 定义如下：

$$\hat{R}_u(t)=\overline{R_u}+\Delta R_{C_u}(t) \tag{5}$$

其中，$\Delta R_{C_u}(t)$ 表示聚类 $C_u$ 中所有的用户对于项目 $t$ 的评分的平均偏差。

$$\Delta R_{C_u}(t)=\sum_{u'\in C_u(t)}(R_{u'}(t)-\overline{R_{u'}})/|C_u(t)| \tag{6}$$

#### 2.2.3 最近邻居选择

首先根据式(7)，选择与 $u_a$ 最相似的聚类作为最近邻居的候选集，然后，根据 Pearson 相关系数，由式(1)计算 $u_a$ 与候选集中每个用户的相似度，找出最相似的 $K$ 个用户构成最近邻居集。

$$\mathrm{sim}(u_a,C)=\frac{\sum_{t\in T_{u_a}\wedge T_C}\Delta R_C(t)\times(R_{u_a}(t)-\overline{R_{u_a}})}{\sqrt{\sum_{t\in T_{u_a}\wedge T_C}(\Delta R_C(t))^2}\times\sqrt{\sum_{t\in T_{u_a}\wedge T_C}(R_{u_a}(t)-\overline{R_{u_a}})^2}} \tag{7}$$

#### 2.2.4 产生推荐

协同过滤是基于最近邻居集的用户评价行为，根据以下公式，为目标用户 $u_a$ 预测其对项目 $t$ 的评分。

$$R_{u_a}(t)=\overline{R_{u_a}}+\frac{\sum_{i=1}^{K}\mathrm{sim}(u_a,u)\times(R_u(t)-\overline{R_u})}{\sum_{i=1}^{K}\mathrm{sim}(u_a,u)} \tag{8}$$

其中 $\mathrm{sim}(u_a,u)$ 表示目标用户与最近邻居集中的用户 $u$ 之间的相似度。

## 3 实验

### 3.1 数据集

本实验的数据来自于 MovieLens，这个数据库包含了 1 682 部电影和 943 个用户。本文从中抽取出评价电影数大于 50 的用户，把满足条件的 563 个用户分为训练集和测试集，其中后 163 个用户为测试集。本文进行 3 次训练，训练集分别表示为 ML_200、ML_300 和 ML_400，大小分别为 200、300 和 400。每次训练前，对训练集进行二分网络的单顶点投影得到用户之间的相似矩阵，为后面的聚类做准备。在此过程中本文不考虑评分值小于 3 的边，排除了两个用户评分行为不一致的情况。训练集用户的谱聚类过程通过 Matlab 实现。

### 3.2 评价指标

本文使用绝对评价误差 MAE(Mean Absolute Error)评价推荐质量：

$$\mathrm{MAE}=\frac{\sum_{u\in T}|R_u(t_j)-\widetilde{R_u}(t_j)|}{|T|} \tag{9}$$

其中，$R_u(t_j)$ 表示训练集中的用户 $u$ 对项目 $t_j$ 实际的评分，$\widetilde{R_u}(t_j)$ 表示用户 $u$ 对项目 $t_j$ 的预测评分。$T$ 是测试集，$|T|$ 表示测试集的大小。MAE 值越小，说明预测的质量越高。

### 3.3 实验结果及分析

本文将聚类数设为 20，分别取最近邻居数为 5、10、20。在实验中，本文将传统的基于 Pearson 相关系数的协同过滤算法作为基线方法进行了比较，并把该方法记为 TCF。对比结果如表 1 所示。

表 1 实验结果数据及其对比

| 训练集 | 方法 | 邻居点 $K=5$ | 邻居点 $K=10$ | 邻居点 $K=20$ |
|---|---|---|---|---|
| ML_200 | TCF | 0.908 | 0.882 | 0.873 |
| | SCBCF | 0.861 | 0.843 | 0.826 |
| ML_300 | TCF | 0.897 | 0.874 | 0.859 |
| | SCBCF | 0.846 | 0.829 | 0.808 |
| ML_400 | TCF | 0.893 | 0.876 | 0.862 |
| | SCBCF | 0.841 | 0.828 | 0.799 |

从表 1 可以看出，协同过滤易受数据稀疏性的影响。本文的方法对训练集的数据进行了平滑处理，从而减轻了这一因素的影响。同时，随着最近邻居数的增加，实验结果也随之改善。这是因为考虑更多相似用户的评分行为，使目标用户的预测评分趋于稳定，从而使得预测值与实际值之间的偏差减小。本文提出的算法在很大程度上缩小了最近邻居候选集的大小，与传统的协同过滤算法相比，算法的伸缩性得到了提高，时间复杂度也进一步降低。

本文考虑将更加高效的谱聚类算法引入到协同过滤推荐中来，实验结果证明本文提出的 SCBCF 算法比传统的协同过滤推荐算法能更好地提高推荐系统的推荐质量。在对用户进行谱聚类时，本文发现聚类结果的各个类之间的用户数并不均衡，这限制了预测能力的进一步提升，因此如何将用户更准确地归类将是未来的研究工作之一。

### 参考文献

[1] SU X，KHOSHGOFTAAR T M.A survey of collaborative filtering techniques[J].Adv.Artif.Intell，2009(1)：421-425.

[2] ZHOU T，REN J，MEDO M，et al.Bipartite network projection and personal recommendation[J].Phys.Rev.E，2007，76(4)：046115-046121.

[3] LUXBURG U.A tutorial on spectral clustering[J].Statistics and Computing，2007，17(4)：395-416.

[4] SARWAR B，KARYPIS G，KONSTAN J，et al.Item-based collaborative ltering recommendation algorithms[C].In www10，Hong Kong，2001.

[5] XUE G R，LIN C，YANG Q，et al.Scalable collaborative filtering using cluster-based smoothing[C].In Proceedings of the ACM SIGIR Conference，Salvador，Brazil，2005：114-121.

(收稿日期：2012-07-17)

作者简介：

张思明，男，1988 年生，硕士研究生，主要研究方向：社会网络链接分析。

游天童，男，1969 年生，教授，博士，硕士生导师，主要研究方向：无线网络。