基于谱聚类群组发现的协同过滤推荐算法

李 贵 陈召新 李征宇 韩子扬 孙 平 孙焕良
(沈阳建筑大学信息与控制工程学院 沈阳 110168)

摘 要 推荐系统中,基于聚类的协同过滤推荐算法利用K-means等算法对用户和物品进行聚类,聚类结果里用户或物品只能属于一个类别,然而在实际应用中,一个用户可以有多种兴趣,一个物品也可以属于多个类别。针对上述问题,提出了一种基于谱聚类群组发现的算法,该算法通过谱聚类和C-means聚类得到用户和物品相似度较高的群组以及用户和物品归属于群组的隶属度矩阵,而且用户或物品可以属于多个群组。通过计算用户在各个群组中对物品的偏好值,并结合用户和物品在群组里相应的隶属度来预测用户对物品最终的偏好值,生成对用户的Top-N推荐结果。实验结果表明,与以往推荐算法相比,本方法在降低了数据稀疏性的同时提高了推荐结果的准确率和召回率。

关键词 推荐系统,协同过滤,谱聚类,C-means算法,群组
中图法分类号 TP301.6 **文献标识码** A

Collaborative Filtering Recommendation Algorithm Based on Spectral Clustering Subgroups Discovering

LI Gui CHEN Zhao-xin LI Zheng-yu HAN Zi-yang SUN Ping SUN Huan-liang
(Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract In many recommendation systems, Collaborative filtering recommendation algorithms based on clustering use some specific algorithms such as the K-means algorithm to cluster the users and items, but the limit is that a user or item can only belong to one category in the clustering result. In practical application, a user may have a variety of interests and an item also belongs to multiple categories. To solve the above problem, this paper put forward a novel algorithm based spectral clustering subgroups discovering and C-means clustering, by which we got the user-item subgroups with a high degree of similarities and the membership matrix of subgroups of users and items, which can belong to multiple subgroups. The purpose of our algorithm is to predict the users' final preference to the items by calculating user's preference to the items in each subgroup and combining the corresponding membership of users and items in their subgroup, and generate the users' top-N recommendation results. Experimental results show that our method reduces the data sparseness and improves the recommendation precision and recall compared with previous recommendation algorithms.

Keywords Recommendation system, Collaborative filtering, Spectral clustering, C-means algorithm, Subgroup

1 引言

推荐系统作为解决互联网信息过载的基本途径,在互联网中有着广泛的应用。推荐系统利用信息发现、数据挖掘等技术在海量的Web信息中为用户推荐文章、电影和音乐等。很多网站已经使用推荐系统为用户提供个性化的服务,如京东商城和亚马逊的产品推荐系统、豆瓣的电影推荐系统、百度MP3音乐推荐系统以及新浪微博的推荐系统等。

协同过滤(Collaborative Filtering)是推荐系统最常用的算法,与基于内容的推荐算法利用用户或物品的属性进行推荐不同,协同过滤算法仅仅利用用户-物品的交互信息进行推荐,如用户对物品的评分等。协同过滤算法通过利用用户的历史行为数据来捕捉用户的爱好,用户的历史行为数据可以是显式的评分数据、用户购买信息等,也可以是隐式的用户浏览记录、用户搜索记录[1]。与基于内容的推荐算法相比,协同过滤算法大大提高了推荐系统的准确率。

传统的协同过滤系统根据目标用户的爱好找到一组与目标用户兴趣相似的用户集合,进而找到这个集合中用户喜好的物品推荐给目标用户。它的基本假设是如果用户有相似的行为就会有相同爱好。但是这个假设并不总是成立的,两个相似的用户可能对这个邻域集合中的物品有着完全不同的兴趣。通常认为,用户的兴趣爱好只是集中在一个或多个主题上,而不是分散在所有的主题上。在推荐系统里,用群组表示相似用户或者物品的集合。本文将相似的用户和物品的集合称为用户-物品群组。

许多基于聚类模型的协同过滤算法利用用户聚类[2]、物品聚类[3]或者用户和物品的联合聚类[4]来得到多个用户物品群组,在这些模型算法中,一个用户或者物品只能属于一个群

本文受国家自然科学基金(61070024),辽宁省自然科学基金(2014020068)资助。
李 贵(1964—),男,博士,教授,主要研究方向为Web数据挖掘与信息集成、分布对象技术、软件工程,E-mail:Ligui21c@sina.com;陈召新(1989—),男,硕士生,主要研究方向为Web数据挖掘和推荐系统。

组。然而，在实际的情况下，一个用户可能对多个主题感兴趣，这样一个用户就可以属于多个群组。例如，一个用户可以喜欢多个主题的电影，一部电影也可能属于多个主题类别。图1展现了两个群组，这里 U_2，U_3 和 I_4，I_5 可以同时属于两个群组。

	I_1	I_2	I_3	I_4	I_5	I_6
U_1				4		5
U_2	5		3		5	4
U_3		4		5	4	
U_4	5		2	2		
U_5	4	1	5		5	

图1 两个重叠的用户-物品群组

针对以往的聚类模型中一个用户或者物品不能同时属于多个群组的问题，本文提出了一种基于谱聚类群组发现的算法，算法记为 SCSD(Spectral Clustering Subgroups discovering)。SCSD 算法通过谱聚类和改进的 C-means 聚类得到用户-物品群组。通过利用现有的推荐算法计算用户在各个群组中对物品的偏好值并结合用户和物品在群组里相应的隶属度来预测用户对物品最终的偏好值。通过两个真实数据集上的实验结果表明，本文提出的算法不仅提高了推荐准确率和召回率，而且有效地降低了数据的稀疏性。

2 相关工作

2.1 基于邻域的协同过滤算法

基于邻域的协同过滤推荐算法是推荐系统中最基本的算法，基于邻域的算法分为两大类：基于用户的协同过滤算法和基于物品的协同过滤算法。

基于用户的协同过滤算法(UserCF)是推荐系统中最早的算法之一，这个算法的出现标志着推荐系统的诞生。基于用户的协同过滤算法通过计算用户和用户之间的相似度，找到和目标用户兴趣最接近的用户集合，进而找到这个集合中其他用户喜欢的物品推荐给目标用户[5]。

基于物品的协同过滤算法(ItemCF)相对于基于用户的协同过滤算法在实际的推荐系统中应用较为广泛，比如著名购物网站亚马逊、京东商城、天猫商城等应用的就是该算法。该算法认为如果喜欢物品 A 的用户也喜欢物品 B，那么就认为物品 A 和 B 具有很大的相似度。基于物品的协同过滤算法首先计算物品之间的相似度，然后为目标用户推荐与他们之前喜欢的物品相似的物品[6,7]。但是用户被限制在只能得到与以往熟悉的内容相类似的物品，不利于挖掘用户潜在的兴趣，作出“跨类型”的推荐。

基于邻域的协同过滤推荐算法需要在整个用户或者物品空间上搜索目标用户的最近邻居，随着用户和物品规模的增加，计算量成线性增长，这降低了推荐系统的效率和系统的实时性。

2.2 基于模型的协同过滤算法

基于模型的协同过滤算法利用训练数据来学习模型进行预测。建模过程通常利用数据挖掘和机器学习技术如贝叶斯模型[8]、矩阵分解模型、聚类模型[2-4]等来实现。本文主要针对基于聚类模型的协同过滤算法进行研究。

聚类是一个对数据集中相似的数据成员进行分类组织的过程[9]。在推荐系统中，聚类用于将具有相同爱好的用户或者相似特征的物品进行分类。聚类的结果可以用作推荐系统的进一步研究。

基于聚类的推荐算法可以划分为基于用户聚类的协同过滤推荐算法(Collaborative Filtering Recommendation Algorithm based on User Clustering)、基于物品聚类的协同过滤推荐算法(Collaborative Filtering Recommendation Algorithm based on Item Clustering)以及基于用户和物品联合聚类的协同过滤推荐算法(Collaborative Filtering Recommendation Algorithm Based on Co-clustering of User and Item)。

Sarwar 等人提出了一种基于用户聚类的推荐算法。该算法根据用户的评分习惯，通过聚类将整个用户集合划分到不同的群组中，然后根据目标用户所属群组来搜索目标用户的最近邻居，大大减少了搜索邻居用户的时间，提高了推荐效率[2]。O'Connor 等人提出了一种基于物品聚类的协同过滤推荐算法，该算法根据用户物品评分数据，通过聚类将物品划分到不同的群组中[3]。Unger 等人利用改进的 K-means 算法和 Gibbs 抽样对用户和物品分别聚类，该算法降低了数据集的稀疏性，提高了推荐质量[4]。

以上 3 种算法无论是对用户或者是对物品聚类，得到的都是用户和物品的严格分组，即每一个用户和物品只能属于一个群组。虽然利用以上 3 种基于聚类的推荐算法可以提高推荐效率，但有时用户的兴趣实际是多样性的，例如一个电影网站，某用户可能对喜剧片感兴趣，但同时他也喜欢动作片，使用以上聚类算法后可能造成只给用户推荐了喜剧类别的电影，而没推荐动作类电影，从而影响了推荐结果的多样性，降低了推荐精度。

3 基于谱聚类群组发现的协同过滤算法

3.1 谱聚类算法

谱聚类(Spectral Clustering)是一种基于图论的聚类算法，源于谱图划分理论。其基本思想是利用样本数据的相似矩阵(拉普拉斯矩阵)的特征向量来对数据点进行聚类。谱聚类与传统的聚类算法如 K-means 等算法相比，聚类结果更优、效率更高[10]。K-means 与 C-means 算法收敛于局部最优解，谱聚类则可以收敛于全局最优解。谱聚类能够识别任意形状的样本空间，相比之下，K-means 的聚类结果一般为超维椭圆体形状。谱聚类的实现主要有 4 个步骤[11,12]：

第 1 步 根据样本数据生成图的邻接矩阵 W。

根据样本数据构造一个图 G，图中的每个顶点对应一个数据点，根据各顶点之间的权重将各顶点连接起来，构成一个无向加权图 $G=(V,W)$，这样对样本数据的聚类问题就转化为图 G 上的图划分问题。图 G 可以用邻接矩阵来表示，记为 W，W 为 $N\times N$ 的对称矩阵，矩阵元素 W_{ij} 表示顶点 v_i 到顶点 v_j 边的权值。

第 2 步 归一化拉普拉斯矩阵。

定义拉普拉斯矩阵 $L=D-W$，其中 D 为图 G 的度矩阵。度矩阵是将邻接矩阵每行或者每列元素相加得到度顶点，然后以所有度顶点为对角元素构成的对角矩阵。$D=diag(D_{11},D_{22},\cdots,D_{nn})$，其中 D_{ii} 为：

$$D_{ii}=\sum_{j=1}^{n}W_{ij} \tag{1}$$

归一化矩阵拉普拉斯矩阵 L：

$$\bar{L}=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}=I-D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{2}$$

其中，$I\in\mathbb{R}^{n\times n}$ 为单位矩阵。

第 3 步　计算矩阵 $\bar{L}$ 最小的 k 个特征值和所对应的特征向量 $s_1,s_2,\cdots,s_k$，构成矩阵 $S=[s_1,s_2,\cdots,s_k]$。

第 4 步　将特征向量通过 K-means 或者 C-means 进行聚类。

将 S 中的每一行看成 k 维空间中的一个向量，并使用 K-means 或者 C-means 算法进行聚类。聚类结果中每一行所属的类别就是原来 G 中的节点亦即最初的数据点分别所属的类别。

3.2　利用谱聚类发现用户-物品群组

3.2.1　构造图 G 和邻接矩阵 W

假设有 n 个用户和 m 个物品，为了得到有意义的用户-物品群组，并且将用户和物品联系到一起，首先构造一个图 G，图 G 中每个用户和物品都是一个顶点，用户和用户之间、用户和物品、物品和物品之间都可以存在边。图 2 展示了包含 4 个用户和物品之间的连通图，可以看出，根据用户物品之间的边权值可以把图分为上下两个重叠部分。

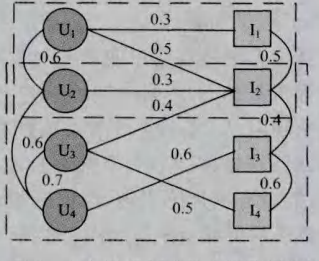


图 2　4 个用户和物品之间连通图

本文使用改进的余弦相似度来度量用户和用户之间、物品和物品之间的相似度。用户和用户之间的边的边权值为用户之间的相似度。设用户 i 和 j 共同评分的物品集合为 I_{ij}，I_i 和 I_j 分别表示用户 i 和用户 j 评分的集合，用户 i 和 j 的相似性 $userSim(i,j)$ 可以表示为：

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

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

物品和物品之间的边权值为物品之间的相似度。设对物品 i 和 j 有共同评分的用户集合为 U_{ij}，U_i 和 U_j 分别表示为物品 i 和物品 j 评分的用户集合，物品 i 和 j 的相似性 $itemSim(i,j)$ 可以表示为：

$$itemSim(i,j)=\frac{\sum\limits_{u\in U_{ij}}(R_{u,i}-\overline{R_u})(R_{u,j}-\overline{R_u})}{\sqrt{\sum\limits_{u\in U_i}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum\limits_{u\in U_j}(R_{u,j}-\overline{R_u})^2}} \tag{4}$$

其中，$R_{u,i}$ 表示用户 u 对物品 i 的评分，$\overline{R_u}$ 表示用户 u 对所有物品的平均评分。

根据文献[13]，用户和物品之间的边权值可以用归一化的用户-物品评分矩阵 S 来表示。S 可以定义为：

$$S=(D^{row})^{-\frac{1}{2}}T(D^{col})^{-\frac{1}{2}} \tag{5}$$

其中，$T\in\mathbb{R}^{n\times m}$ 为用户-物品评分矩阵，$D^{row}\in\mathbb{R}^{n\times n}$ 和 $D^{col}\in\mathbb{R}^{m\times m}$ 都是对角矩阵，且 $D_{ii}^{row}=\sum\limits_{j=1}^{m}T_{ij}$，$D_{jj}^{col}=\sum\limits_{i=1}^{n}T_{ij}$。

这样，可以构建图 G 的邻接矩阵 W。

$$W=\begin{bmatrix}Q & S\\ S^{\mathrm{T}} & R\end{bmatrix} \tag{6}$$

其中，$Q\in\mathbb{R}^{n\times n}$ 为用户之间的相似度矩阵，$R\in\mathbb{R}^{m\times m}$ 为物品之间的相似度矩阵，$S\in\mathbb{R}^{n\times m}$ 为用户评分的归一化矩阵。

3.2.2　求出矩阵 $\bar{L}$ 最小的 k 个特征值和对应特征向量

由式(4)可以得到归一化的拉普拉斯矩阵：

$$\bar{L}=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}=I-D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{7}$$

其中，$L=D-W$，$D\in\mathbb{R}^{(n+m)\times(n+m)}$ 为对角矩阵，且 $D_{ii}=\sum\limits_{j=1}^{m+n}W_{ij}$，$I\in\mathbb{R}^{(n+m)\times(n+m)}$ 为单位矩阵。

由特征值问题 $\bar{L}X=\lambda X$ 求出矩阵 $\bar{L}$ 最小的 k 个特征值和对应的特征向量，组成矩阵 $M\in\mathbb{R}^{(n+m)\times k}$，将矩阵 M 进行转置，得到 $M^{\mathrm{T}}=[x_1,\cdots,x_{n+m}]$。

3.2.3　通过 C-means 算法得到用户-物品群组

本文使用分块矩阵 $P\in[0,1]^{(n+m)\times k}$ 来表示聚类结果，P_{ij} 的大小表示第 i 个样本属于第 j 个群组的权重大小，每个样本可以属于的群组数为 k，即 P 的每一行只能有 k 个非负元素，且 P 的每行之和都为 1。分块矩阵 P 可以写为：

$$P=\begin{bmatrix}U\\ I\end{bmatrix} \tag{8}$$

其中，矩阵 $U\in[0,1]^{n\times k}$ 是用户隶属度矩阵，U_{ij} 表示第 i 个用户归属于第 j 个群组的权重大小。矩阵 $I\in[0,1]^{m\times k}$ 表示物品隶属度矩阵，I_{ij} 表示第 i 个物品归属于第 j 个群组的大小。

为了使一个用户或者物品可以属于多个群组，本文使用模糊 C-means 聚类算法[14]来对 $M^{\mathrm{T}}=[x_1,\cdots,x_{n+m}]$ 中的列向量 $x_1,\cdots,x_{n+m}$ 进行聚类，聚类的结果中 $x_1,\cdots,x_{n+m}$ 每一列所属的类别就是图 G 中对应的用户或物品节点分别所属类别。该算法是一个迭代最优化算法，最小化如下最优化函数：

$$J_m(P,V)=\sum_{i=1}^{m+n}\sum_{j=1}^{c}(P_{ij})^l d(x_i,v_j)^2 \tag{9}$$

其中，x_i 是样本数据，v_j 是群组 j 的中心。函数 d 是一个可以预先定义的距离函数，参数 l 是分块结果中控制模糊性的加权指数。在本文的方法中 d 是欧几里得距离。

由于模糊 C-means 聚类算法对初始值非常敏感，聚类结果受初始值的影响很大。一般认为，比较合适的聚类中心出现在样本点比较密集的地方[15]。为了得到合适的初始点，本文使用密度函数法来确定初始点。

定义样本点 x_i 处的密度函数：

$$D_i^{(0)}=\sum_{c=1}^{n+m}\frac{1}{1+f_d\parallel x_i-x_c\parallel^2} \tag{10}$$

其中，$f_d=\dfrac{16((n+m)(n+m-1))}{\parallel x_i-x_c\parallel}$，$n+m$ 为样本的个数。令 $D_1=\max\{D_i^{(0)},i=1,\cdots,n+m\}$，对应的 v_1 取第一个初始聚类中心，后续初始聚类的中心的密度调整关系如下：

$$D_i^{(c)}=D_i^{(c-1)}-D_c\frac{1}{1+f_d\parallel x_i-v_c\parallel} \tag{11}$$

其中，$D_c=\max\{D_i^{(c-1)},i=1,\cdots,n+m\}$，$c=1,2\cdots,k-1$，对应的样本点 v_k 为第 k 个初始聚类中心。

对于所有的 $i=1,\cdots,m+n$ 和 $j=1,\cdots,k$，在每次迭代中，我们按照如下公式更新 P 和 V。

$$P_{ij}=\frac{(d(x_i,v_j))^{2/(1-l)}}{\sum\limits_{c=1}^{k}(d(x_i,v_c))^{2/(1-l)}} \tag{12}$$

$$v_j=[\sum_{i=1}^{m+n}P_{ij}^l x_i]/[\sum_{i=1}^{m+n}P_{ij}^l] \tag{13}$$

如果目标函数的值在两个迭代中小于一个阈值 ε，则迭代停止。在本文的实验中，为了取得最优值，根据参考文献[17]，$l=2,\varepsilon=e^{-5}$。这样对于 P 的每一行，仅仅有 k 个最大的元素被保留，且每一行的和为 1。

3.3 利用用户-物品群组进行推荐

通过以上算法得到用户-物品归属矩阵 P，从而就得到了用户-物品群组。这些用户-物品群组可以看成原始用户-物品矩阵 T 的子矩阵。这样不用对现有的推荐算法进行任何修改就可以通过输入群组的用户-物品子矩阵来得到推荐结果。

为了得到准确的推荐结果，需要合并从所有群组中得到的预测结果。本文通过用户对物品在群组中预测评分并结合相应的隶属度来预测用户的最终评分。

一个用户和物品可以属于一个或者多个群组也可以不属于任何群组。定义 $Pre(i,j,k)$ 是用户 i 对物品 j 通过协同过滤算法在群组 k 中的预测评分，F_{ij} 是用户 i 对物品 j 的最后的预测评分。当 i 和 j 不属于任何群组时，$F_{ij}=0$，当 i 和 j 属于一个或者多个群组时，在本文的方法中，最终的预测评分为各个群组预测评分的加权平均[16]。

用户 i 和物品 j 属于群组 k 的隶属度可以表示为：

$$prob_k=\frac{P_{ik}P_{jk}}{\sum_{k'=1}^{c}P_{ik'}P_{jk'}} \tag{14}$$

其中，P_{ik} 表示用户 i 属于群组 k 的隶属度，P_{jk} 表示物品 j 属于群组 k 的隶属度，P 是由上文中得到的隶属度矩阵。

用户 i 对物品 j 的最后预测评分 F_{ij} 可以表示为：

$$F_{ij}=\frac{\sum_{k=1}^{c}Pre(i,j,k)\times prob_k}{\sum_{k=1}^{c}prob_k} \tag{15}$$

这样就得到了用户所有物品的评分，取前 N 个最大的评分生成对用户的 Top-N 推荐结果。

4 实验结果及分析

4.1 数据集及度量标准

本文采用 MovieLens 100k 和房谱网(www.house-book.com.cn)真实数据集(见表 1)。MovieLens 数据集被广泛地应用验证推荐算法的推荐性能。MovieLens 100k 数据集包含 943 多个用户对 1682 部电影的评分的 10 万条评分。该数据集是一个评分数据集，用户可以给电影评 5 个不同等级的分数(1～5 分)。为了验证算法在真实的推荐系统中的表现，本文使用了从房谱网采集的真实数据集。房谱网的真实数据集是来自 4200 个用户对 1203 个楼盘的 8 万多条评分记录，该数据集和 MovieLens 一样是一个评分数据集，用户可以给每个楼盘不同等级的评分(1～5 分)。

表 1 Movielens 和房谱网数据集基本统计信息

数据集	用户数	物品数	评分数量
Movielens 100k	943	1682	100000
HouseBook	4200	1203	85894

许多以前的推荐系统研究都是基于用户评分数据的评分预测，评分预测的预测准确度一般是通过平均绝对误差 MAE (Mean Absolute Error)与均方根误差 RMSE (Root Mean Square Error)来计算。对于一个推荐系统，推荐的目的是找到用户最感兴趣的物品，而不是预测使用过物品后会给物品什么样的评分，因而 Top-N 推荐更符合实际的需求。Top-N 一般通过准确率(Precision)和召回率(Recall)来度量。$R(u)$ 是根据用户在训练集上的行为给用户做出的推荐列表，$T(u)$ 是用户在测试集上的行为列表。

推荐系统的召回率可以定义为：

$$Recall=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{16}$$

推荐系统的准确率可以定义为：

$$Precision=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \tag{17}$$

4.2 实验设置

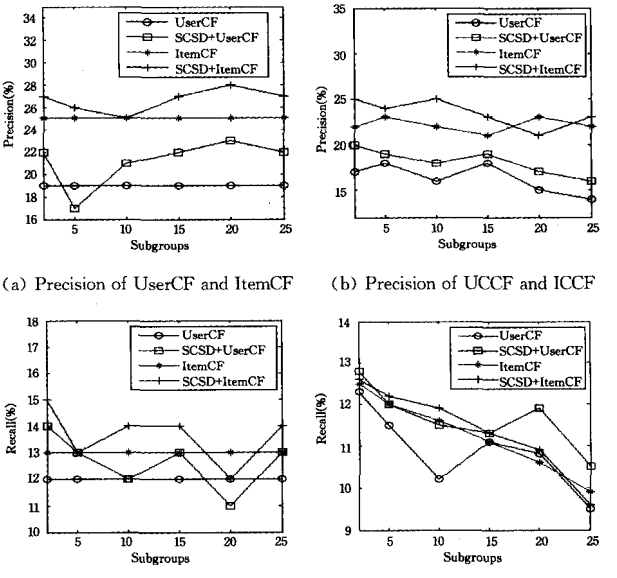
将原数据集随机分为训练集与测试集，其中 70%是训练集，30%是测试集。利用训练集生成模型，针对测试集中的每个物品进行评分预测。

一个物品成为某一个用户的候选推荐物品当且仅当它们都属于一个或多个用户群。但有时候，某些群组由于聚类不均衡，只有很少的元素，例如少于 10 个。在这种情况下，可以去除这些小的用户群或者为这些用户添加一些流行的物品，后者是一个真实的推荐系统经常采用的策略，本文实验亦遵从这个策略。

由于 SCSD 算法可以直接使用现有推荐算法得到推荐结果，本文将使用 SCSD 算法的协同过滤推荐算法和未使用 SCSD 算法的协同过滤推荐算法的推荐结果进行比较。本文实验使用的协同过滤推荐算法为：基于用户的协同过滤算法(UserCF)[5]、基于物品的协同过滤算法(ItemCF)[6]、基于用户聚类的协同过滤算法(UCCF)[2]以及基于物品聚类的协同过滤算法(ICCF)[3]。

4.3 实验结果及分析

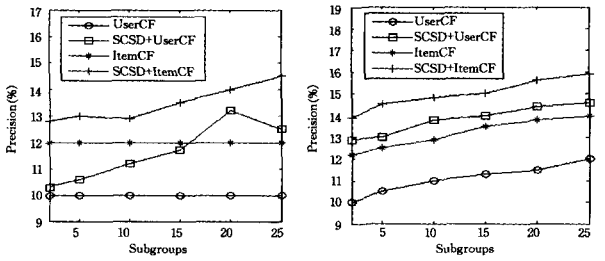
图 3、图 4 为应用 SCSD 算法的推荐算法和原推荐算法在 MovieLens 100k 以及房谱网数据集上准确率和召回率的比较。表 2 为 MovieLens 100k 数据集和房谱网数据集在不同群组数量下的数据稀疏性的比较。本文中数据的稀疏性用矩阵中评分为零的元素所占的百分比来表示。



(a) Precision of UserCF and ItemCF (b) Precision of UCCF and ICCF

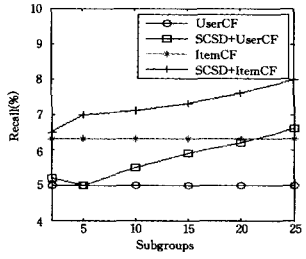
(c) Recall of UserCF and ItemCF (d) Recall of UCCF and ICCF

图 3 各算法基于 MovieLens 100k 数据集准确率和召回率的比较

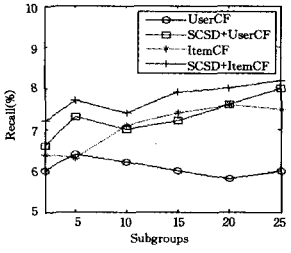


(a) Precision of UserCF and ItemCF

(b) Precision of UCCF and ICCF



(c) Recall of UserCF and ItemCF



(d) Recall of UCCF and ICCF

图 4　各算法基于房谱网数据集准确率和召回率比较

表 2　Movielens 和房谱网数据集在不同群组数下稀疏性比较

	Movielens 100k	房谱网数据集
原始矩阵	95%	98.3%
5 个群组	94.2%	94.2%
10 个群组	93.3%	92.1%
15 个群组	82.3%	86.5%
20 个群组	76.2%	77.6%
25 个群组	73.2%	72.6%

从图 3、图 4 和表 2 可以看出，与原协同过滤算法相比，应用 SCSD 的协同过滤算法不仅提高了推荐结果的准确率和召回率，而且有效地降低了数据的稀疏性。特别是在房谱网真实数据集上，该数据集和 Movielens 数据集相比更为稀疏，采用 SCSD 算法后，数据的稀疏性明显降低，并且推荐结果的准确率和召回率有着较大的提升。

结束语　本文通过分析传统的基于聚类的协同过滤算法聚类结果不能属于多个类别的问题，提出了基于谱聚类和改进的 C-means 聚类的协同过滤算法。通过谱聚类和 C-means 模糊聚类算法得到用户-物品群组，然后通过用户对物品在各个群组中预测评分并结合相应的隶属度来预测用户的最终评分。通过两个数据集上的实验结果表明，该方法不仅提高了数据推荐的准确率和召回率，而且有效地降低了数据了稀疏性。本文使用的数据只有用户-物品的评分数据，用户的浏览、搜索的等隐式数据并没有加入，结合这些隐式的数据更能理解用户的行为，以进行更准确的推荐。同时本文聚类数目为自己定义，如何利用聚类本身特征自动确定聚类数目也是今后需要研究的内容。

参 考 文 献

[1] 项亮，陈义，王益. 推荐系统实践[M]. 河北：人民邮电出版社，2012：39-43

[2] Sarwar B, Karypis G, Konstan J, et al. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering[C]// Proceedings of the Fifth International Conference on Computer and Information Technology. 2002: 158-167

[3] O'Connorand M, Herlocker. Clustering items for collaborative filtering[C] // Proceedings of the ACM SIGIR Workshop on Recommender Systems. 1999

[4] Breese J, Heckerman D, Kadie C, et al. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. 1998: 43-52

[5] Redpath J, Glass D H, McClean S. User-based Collaborative Filtering: Sparsity and Performance[C]// Proceeding of the 2010 Conference on STAIRS. 2010: 287-310

[6] Barbieri N, Manco G. An analysis of probabilistic methods for top-N recommendation in collaborative filtering[C] // Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases. Athens, Greece, 2011: 172-187

[7] Cheng Guang-hua, Gong Song-Jie. An Efficient Collaborative Filtering Algorithm with Item Hierarchy[C]// Proceedings of the 2008 Second International Symposium on Intelligent Information Technology Application. 2008: 28-31

[8] Heckerman D, Chickering D, Meek C, et al. Dependency networks for inference, collaborative filtering, and data visualization [C]// The Journal of Machine Learning Research. 2001, 1: 49-75

[9] Hruschka E R, Campello R J G B, Freitas A A. A survey of evolutionary algorithms for clustering[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2009, 2(39): 133-155

[10] Shi J, Malik J. Normalized Cuts and Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[11] Wang S, Siskind J M. Image segmentation with ratio cut[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2003, 25(6): 675-690

[12] Meila M, Xu L. Multiway cuts and spectral clustering [R]. Washington: University of Washington, 2003

[13] Xu Bin, Bu Jia-jun, Chen Chun, et al. An exploration of improving collaborative recommender systems via user-item subgroups[C]// Proceedings of the 21st International Conference on World Wide Web, WWW'2012. 2012: 21-30

[14] Leski J. Towards a robust fuzzy clustering[J]. Fuzzy Sets and Systems, 2003, 137(2): 215-233

[15] 裴继红，范九伦，谢维信. 聚类中心的初始化方法[J]. 电子科学会刊. 1999, 21(3): 320-325

[16] 吴湖，王永吉，王哲，等. 两阶段联合聚类协同过滤算法[J]. 软件学报，2010, 5(21): 1042-1054

[17] 高新波，裴继红，谢维信. 模糊 c-均值聚类算法中加权指 m 的研究[J]. 电子学报，2000(4): 80-83