计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20160175
53(8):1664-1672,2016

融合社区结构和兴趣聚类的协同过滤推荐算法

郭弘毅　刘功申　苏　波　孟　魁
(上海交通大学电子信息与电气工程学院　上海　200240)
(king-guo@sjtu.edu.cn)

Collaborative Filtering Recommendation Algorithm Combining Community Structure and Interest Clusters

Guo Hongyi, Liu Gongshen, Su Bo, and Meng Kui
(*Shool of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai* 200240)

Abstract　Traditional collaborative filtering recommendation algorithms suffer from data sparsity, which results in poor recommendation accuracy. Social connections among users can reflect their interactions, which can be mixed into recommendation algorithms to improve the accuracy. Only straightforward social connections have been used by most current social recommendation algorithms, while users' latent interest and cluster information haven't been considered. In response to these circumstances, this paper proposes a collaborative filtering recommendation algorithm combining community structure and interest clusters. Firstly, overlapping community detection algorithm is used to detect the community structure existed in user social network, thus users in the same community have certain common characteristics. Meanwhile, we design a customized fuzzy clustering algorithm to discover users' interest clusters, which uses item-category relationship and users' activity history as input. Users in the same cluster are similar in generalized interest. We quantify users' preference for each social community and interest cluster they belong to respectively. Then, we combine this two types of user group information into matrix factorization model by adding a clustering-based regularization term to improve the objective function. Experiments conducted on the Yelp dataset show that, in comparison to other methods including both traditional and social recommendation algorithms, our approach gets better recommendation results in accuracy.

Key words　overlapping community; interest cluster; recommendation algorithm; collaborative filtering; matrix factorization

摘　要　传统的协同过滤推荐算法受限于数据稀疏性问题,导致推荐结果较差.用户的社交关系信息能够体现用户之间的相互影响,将其用于推荐算法能够提高推荐结果的准确度,目前的社交化推荐算法大多只考虑了用户的直接社交关系,没有利用到潜在的用户兴趣偏好信息以及群体聚类信息.针对上述情况,提出一种融合社区结构和兴趣聚类的协同过滤推荐算法.首先通过重叠社区发现算法挖掘用户社交网络中存在的社区结构,同时利用项目所属类别信息,设计模糊聚类算法挖掘用户兴趣偏好层面的聚类信息.然后将2种聚类信息融合到矩阵分解模型的优化分解过程中.在Yelp数据集上进行了新算法与

收稿日期:2016-03-17;修回日期:2015-05-24
基金项目:国家"九七三"重点基础研究发展计划基金项目(2013CB329603);国家自然科学基金项目(61472248,61431008)
This work was supported by the National Basic Research Program of China (973 Program) (2013CB329603) and the National Natural Science Foundation of China (61472248,61431008).

其他算法的对比实验,结果表明,该算法能够有效提高推荐结果的准确度.

关键词 重叠社区;兴趣聚类;推荐算法;协同过滤;矩阵分解

中图法分类号 TP311; TP181

近年来,随着信息技术和互联网的飞速发展,人们逐渐从信息匮乏的时代走入了信息过载的时代,如何从海量数据中筛选出有价值的信息是信息消费者和信息提供者都要面对的挑战.推荐系统作为联系用户和信息的工具由此应运而生,它能够使得信息消费者获取对自己感兴趣的信息,同时使得信息提供者能够有针对性地向目标用户投放信息,实现两者的共赢.目前,推荐系统已经不同程度地运用到了多个互联网领域中[1].

推荐算法是推荐系统的关键部分.其中,协同过滤推荐算法是目前应用最为成功的推荐技术之一[2].其主要思想是利用目标用户的消费记录,基于消费行为或者评分相似的用户具有相似消费偏好的假设,找到与目标用户偏好相似的用户集合,根据用户集合的喜好给目标用户推荐其可能感兴趣的项目.相较于基于内容的推荐算法,尽管协同过滤推荐算法不依赖于项目的特征信息,不受限于内容分析技术的局限,但是受限于数据稀疏性问题[3].互联网规模的急速扩张带来了用户数量和项目数量的急剧增长,用户-项目评分信息稀疏,用户间共同消费的项目很少,直接导致推荐结果的准确度下降.

随着 Web2.0 的迅速发展,互联网用户能够扮演愈发活跃的角色,除了常规的购买和评分行为,还能够与信赖的或志同道合的用户建立信任关系.用户的消费行为也不再仅仅是其个人的兴趣偏好的体现,而且在一定程度上也受到与其具有社交关系的用户的影响.社交网络分析的研究表明,网络社区中,受到社交因素的影响,有社交关联的用户往往会体现出相似的兴趣爱好和行为特征[4].因此,融入用户社会属性的社交化推荐系统成为近年来推荐系统领域的研究热点.其中传统的社交化推荐算法采用了基于信任的模型,利用了用户间的直接信任关系,然而随着互联网规模的不断扩大,用户间的直接信任关系不可避免地出现数据稀疏性的问题.此外,基于信任模型的社交化推荐算法的基本假设是用户的兴趣偏好与其所信任的用户相似或者受到这些用户的影响[5].然而在现实生活中,用户的兴趣偏好是多方面的,其信任的个体间的兴趣偏好也存在差异,单一的直接社交关系并不能刻画出针对不同领域的项目时用户与好友的兴趣偏好的差异性.

基于上述原因,本文提出一种融合社区结构和兴趣聚类的协同过滤推荐算法.通过重叠社区发现算法挖掘用户的社交网络结构中蕴含的社区信息,避免了使用直接社交关系引起的数据稀疏性问题,量化不同用户在社区中的影响力的差异.此外,考虑到同一社区中用户群体的兴趣偏好的差异,利用项目类别信息,挖掘用户兴趣偏好层面的聚类信息.将 2 种聚类信息融合到基于矩阵分解的协同过滤算法中,通过对矩阵分解过程中的隐式特征向量进行约束来优化目标函数.

1 相关工作

传统的协同过滤推荐算法分成基于内存的方法和基于模型的方法 2 类[6].近年来,基于矩阵分解模型的协同过滤推荐算法作为基于模型的方法中的一个分支被广泛应用到了推荐系统中.它能够把高维的用户-项目评分矩阵转化成表示用户和项目隐式特征向量的低维矩阵乘积的形式,实现高维数据的降维,在缓解数据稀疏性造成的推荐结果准确度下降问题方面有非常好的效果.文献[7]首先提出了矩阵分解技术在推荐系统领域的应用,文献[8]提出了概率矩阵分解模型,从条件概率最优的角度进行矩阵的概率优化分解,得到了相同的矩阵分解模型.

在社交化推荐算法中,矩阵分解技术融合了用户的各种社会属性信息,基于用户与所信任的用户群体具有相似的兴趣偏好或者受其影响的假设,通过在矩阵优化分解的过程中添加相关约束得到更优的用户隐式特征向量.文献[9]提出了 SoRec 模型,它是一种社会谱正则化的变形方法,把矩阵分解技术同样作用在信任矩阵上,用户隐式特征向量同时从用户-项目评分矩阵和信任矩阵的优化分解过程中得到.文献[10]提出了 STE 模型,把用户-项目评分矩阵中的项看作是用户个人偏好以及用户信任好友喜好的组合,在优化分解过程中将用户对项目的评分和用户的朋友对项目的评分加权平均,使得推荐结果拥有了可解释性.文献[11]提出了 SocialMF 模型,假设用户的隐式特征向量是由其朋友的隐式

特征向量决定的,在优化分解过程中引入了信任传播的概念.这些社交化推荐算法只利用了用户的直接社交关系,当直接社交关系稀疏时会导致推荐结果不理想.文献[12]首次在社交化推荐算法中引入了重叠社区发现算法,关注对目标函数中的正则项的约束.提出了2种模型旨在减小用户与其所在社区其他用户的偏好的差异.文献[13]在SocialMF模型的基础上进行改进,区分了对于不同项目类别,用户对不同好友的信任度的差异.算法直接根据项目类别划分用户的好友,因此可能会进一步加剧数据稀疏性问题.文献[14]考虑到信任多样性的特点.对用户信任关系和用户兴趣进行建模识别出目标用户信任且兴趣接近的用户,改进SocialMF模型.但算法仍可能受限于直接社交关系稀疏问题.

2 融合社区结构和兴趣聚类的推荐算法

本文提出的融合社区结构和兴趣聚类的协同过滤推荐算法的流程如图1所示.1)利用重叠社区发现算法挖掘用户社交网络中存在的社区结构,得到基于网络结构的用户集合;2)利用针对数据特性改进的模糊C均值聚类算法,根据项目类别信息和用户行为记录得到基于兴趣偏好相似度的用户集合,分别量化目标用户对其所属的2类不同用户集合的感兴趣程度;3)将2种用户聚类信息融合到矩阵分解模型的优化分解过程中,通过在目标函数中引入新的正则项试图得到更优的分解结果,最终获得用户对项目的预测评分.

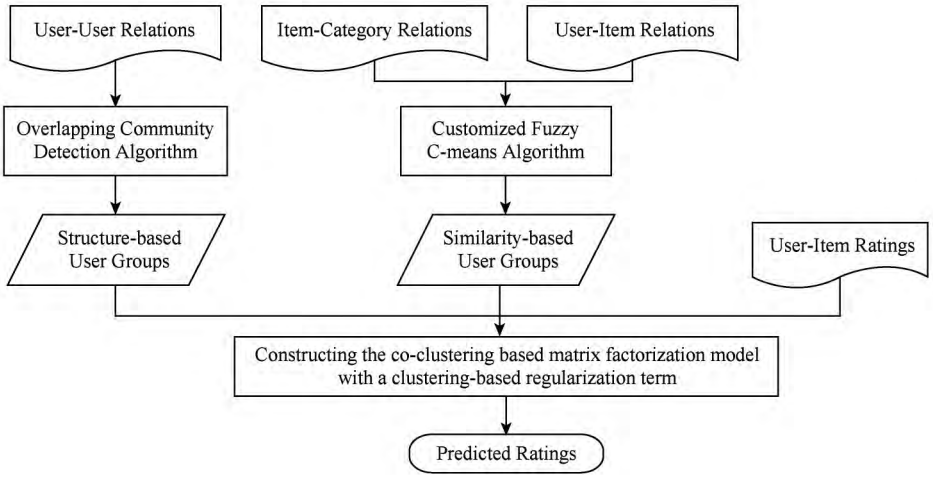

Fig. 1 The flow chart of co-clustering recommendation algorithm.

图1 融合双重聚类的推荐算法流程图

2.1 准备工作

设$U=\{u_1,u_2,\cdots,u_m\}$表示推荐系统中所有用户的集合,$V=\{v_1,v_2,\cdots,v_n\}$表示推荐系统中所有项目的集合,$C=\{c_1,c_2,\cdots,c_q\}$表示推荐系统中所有项目类别的集合,其中m,n,q分别表示用户总数、项目总数、类别总数.$\boldsymbol{R}=(R_{ij})^{m\times n}$表示用户-项目评分矩阵,其中$R_{ij}\in\{1,2,3,4,5\}$表示用户$u_i$对项目$v_j$的评分.$\boldsymbol{T}=(T_{ij})^{m\times m}$,$T_{ij}\in\{0,1\}$表示用户的社交关系矩阵,$T_{ij}=1$表示用户$u_i$与用户$u_j$之间存在好友关系.本文采取双向确认的用户的社交关系,因此矩阵$\boldsymbol{T}$为对称矩阵.

2.2 基于社区结构的聚类

推荐系统中的用户集合以及用户间的社交关系构成了庞大的社交网络,用户往往与其直接好友具有相似的兴趣偏好或者受其影响,一些研究[9-10,15]即根据这个假设在传统的协同过推荐算法加入用户的社交关系信息进行优化.然而在大型社交网络中普遍存在着长尾效应[16],即社交关系多的用户占总量的少数,而绝大多数用户只有很少的社交关系.因此,有必要在网络中挖掘其他有价值信息.社交网络中总是存在社区结构,同一个社区内的用户具有某些相同的特性,如地理位置相近、行业领域相同、关注的内容主题相近等.社区内的其他用户或多或少地会对用户的选择产生影响.社交网络中的用户往往同属于多个社区,比如用户与其所在城市相同的用户属于一个社区,与喜爱科幻类电影的用户同属于另外一个社区.这些重叠社区信息体现了用户的不同特性.

对于社交网路中的重叠社区发现的研究是近年来社区发现领域的研究热点,本文直接采用其中效

果突出的重叠社区发现算法来划分推荐系统中的用户社交网络. BIGCLAM 算法是一种适用于大型网络的重叠社区发现算法[17],它基于社区间重叠部分中的节点紧密连接的假设,在非负矩阵分解模型的基础上进行改进. 文献[12]的实验比对结果表明使用 BIGCLAM 算法得到的社区信息作为社交化推荐算法的约束条件能够取得较好的推荐结果. 因此,本文选择 BIGCLAM 算法划分用户社交网络中的重叠社区.

显然,用户对于其所在的不同社区的感兴趣程度存在差异,文献[12]中设定某个社区内所有用户对应在用户-评分矩阵内的用户评分向量的平均值作为社区评分向量,计算属于该社区的某用户对应的用户评分向量与该社区评分向量的相似度作为该用户对该社区的感兴趣程度. 然而社区中的每个用户对该社区做出的贡献是不同的. 相对于处于社区结构边缘的用户,在社区中拥有更多与其有直接社交关系的好友的用户更能够代表这个社区. 基于该假设考虑社区中所有用户的评分向量和社区好友数量,从而获取带权重的社区评分向量:

$$\boldsymbol{Com}(i)=\frac{\sum\limits_{g\in\Omega(i)}|friend_i(g)|\boldsymbol{U}_g^{\boldsymbol{R}}}{\sum\limits_{g\in\Omega(i)}|friend_i(g)|},\tag{1}$$

其中,$\Omega(i)$表示社区 i 中所有用户的集合,$friend_i(g)$ 表示社区 i 中与用户 u_g 有直接社交关系的好友集合,$\boldsymbol{U}_g^{\boldsymbol{R}}$ 为用户评分向量. 由式(1)可以看出,在社区中拥有更多与其有直接社交关系的好友的用户对社区评分向量的贡献度更大. 接着,通过计算社区评分向量和用户评分向量的皮尔逊相关系数(Pearson correlation coefficient)得到社区评分向量和用户评分向量的相似度:

$$Sim(i,j)=\frac{\sum\limits_{f\in\Lambda_{ij}}(U_{if}^{\boldsymbol{R}}-\overline{U}_i^{\boldsymbol{R}})(U_{jf}^{\boldsymbol{R}}-\overline{U}_j^{\boldsymbol{R}})}{\sqrt{\sum\limits_{f\in\Lambda_{ij}}(U_{if}^{\boldsymbol{R}}-\overline{U}_i^{\boldsymbol{R}})^2}\sqrt{\sum\limits_{f\in\Lambda_{ij}}(U_{jf}^{\boldsymbol{R}}-\overline{U}_j^{\boldsymbol{R}})^2}},\tag{2}$$

其中,Λ_{ij} 表示用户评分向量 $\boldsymbol{U}_i^{\boldsymbol{R}}$ 和用户评分向量 $\boldsymbol{U}_j^{\boldsymbol{R}}$ 中元素均不为零的位置集合. 皮尔逊相关系数的输出值范围为[-1,1],使用 $f(x)=(x+1)/2$ 将输出值映射到[0,1]之间.

由此,我们得到了基于社交网络结构的用户社区信息,用户评分向量与所属社区的社区评分向量的相似度即表示用户对该社区的感兴趣程度. 处在同一社区中的用户具有相同的特性或相互影响.

2.3 基于兴趣偏好的聚类

重叠社区发现算法将用户集合根据其社交网络结构进行划分,属于同一社区内的用户存在相同的特性或相互影响. 然而同一社区内的用户依然可能存在不同的兴趣偏好. 如喜爱科幻电影的用户被划分入一个社区,但他们对于音乐、游戏、饮食的偏好却存在很大差异,因此有必要对同一社交网络社区中的用户进行进一步的划分. 基于以上原因,提出了基于兴趣偏好的模糊聚类算法,该算法利用用户的行为记录以及项目所属的类别,寻找与目标用户在泛化层面的兴趣偏好相似的用户集合.

1) 定义用户类别偏好向量

用户评分过的项目可能属于不同的类别,用户对某一类别中的项目的评分数量占的比例与该用户对该类别感兴趣的程度成正比. 用户 u_i 评分过的所有项目的所属类别的分布向量可描述如下:

$$\boldsymbol{Pre}(i)=\left(\frac{|P_i^{c_1}|}{|P_i|},\frac{|P_i^{c_2}|}{|P_i|},\cdots,\frac{|P_i^{c_q}|}{|P_i|}\right),\tag{3}$$

其中,P_i 表示用户 u_i 评分过的项目集合,$P_i^{c_k}$ 表示用户 u_i 评分过的属于类别 c_k 的项目集合.

2) 定义聚类目标函数

模糊 C 均值聚类(fuzzy C-means)[18]作为模糊聚类分析中的一个比较成熟的算法,在各个领域都有着广泛的应用. 其通过优化目标函数来获取每个样本点属于某个类簇的隶属度,即样本点可以同时属于多个类簇,符合推荐系统中用户兴趣偏好的性质. 模糊 C 均值聚类在计算样本点向量和类簇中心点向量的相似度时通常采用的是闵可夫斯基距离. 即

$$\begin{aligned}&D(\boldsymbol{X},\boldsymbol{Y})=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p},\\&\boldsymbol{X}=(x_1,x_2,\cdots,x_n)\in\mathrm{R},\\&\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)\in\mathrm{R}.\end{aligned}\tag{4}$$

当 $p=1$ 时,$D(\boldsymbol{X},\boldsymbol{Y})$为曼哈顿距离,当 $p=2$ 时为欧几里得距离. 然而推荐系统中的用户-项目评分矩阵数据稀疏,用户大多只对某几个类别的项目产生过评分行为,导致用户类别偏好向量也存在稀疏性问题,直接使用闵可夫斯基距离公式计算得出的相似度可能会影响聚类的效果.

基于以上原因,需要设计能够处理用户类别偏好向量存在的数据稀疏性问题的相似度测量方式. 可描述如下:

$$D_{\mathrm{sparse}}(\boldsymbol{Pre}(i),\boldsymbol{g}_j)=\frac{1}{\delta_i}\sum_{k=1}^{q}\delta_{i,k}(Pre(i)_k-g_{j,k}),\tag{5}$$

其中，$\delta_{i,k}=\begin{cases}1, & Pre(i)_k\neq 0\\ 0, & Pre(i)_k=0\end{cases}$，$\delta_i=\sum_{k=1}^{q}\delta_{i,k}$.

式(5)表示样本点向量与类簇中心向量的相似度. 我们假设类簇的数量为 l. 式(5)中 $\boldsymbol{g}_j$ 表示某个类簇中心点 $j(j=1,2,\cdots,l)$的向量. $g_{j,k}$ 为 $\boldsymbol{g}_j$ 中的第 k 个元素. D_{sparse} 在用户类别偏好向量中选取非零元素即用户产生过评分行为的类别元素计算相似度.

有了相似度表达式，我们接着定义目标函数的表达式：

$$Obj(\boldsymbol{B},\boldsymbol{G})=\sum_{i=1}^{m}\sum_{j=1}^{l}\mu_{ij}^{\theta}D_{\text{sparse}}(\boldsymbol{Pre}(i),\boldsymbol{g}_j),\quad(6)$$

其中，$\mu_{ij}\in[0,1]$表示用户 u_i 对类簇 Ψ_j 的隶属程度，μ_{ij} 满足 $\sum_{j=1}^{l}\mu_{ij}=1$. $\boldsymbol{B}=(\mu_{ij})^{m\times l}$表示用户类簇隶属度矩阵. $\boldsymbol{G}=(\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_l)^{\mathrm{T}}$ 表示类簇中心矩阵，$\theta\in[0,\infty)$表示聚类模糊程度，通常置为 2[19].

3) 聚类算法过程

基于兴趣偏好的模糊聚类算法通过迭代运算，更新用户类簇隶属度矩阵 $\boldsymbol{B}$ 和类簇中心矩阵 $\boldsymbol{G}$ 的值，逐步减小目标函数的误差值，当目标函数的误差值收敛至预设阈值时迭代终止. 具体过程如下：

① 随机初始化用户类簇隶属度矩阵 $\boldsymbol{U}$，要求满足 $\sum_{j=1}^{l}\mu_{ij}=1$. 选择合适的类簇数量 l，聚类模糊度 θ，判定收敛阈值 $\varepsilon\in(0,1)$，最大迭代次数 $t_{\max}$.

② 更新类簇中心矩阵 $\boldsymbol{G}$

$$\boldsymbol{g}_j=\sum_{i=1}^{m}\mu_{ij}^{\theta}\boldsymbol{Pre}(i)\Big/\sum_{i=1}^{m}\mu_{ij}^{\theta},\quad(7)$$

其中，$j=1,2,\cdots,l$.

③ 更新用户类簇隶属度矩阵 $\boldsymbol{B}$

$$\mu_{ij}=1\Big/\sum_{k=1}^{l}\left(\frac{D_{\text{sparse}}(\boldsymbol{Pre}(i),\boldsymbol{g}_j)}{D_{\text{sparse}}(\boldsymbol{Pre}(i),\boldsymbol{g}_k)}\right)^{\frac{1}{\theta-1}},\quad(8)$$

其中，$i=1,2,\cdots,m$; $j=1,2,\cdots,l$.

④ 如果 $\|\boldsymbol{B}^{(t+1)}-\boldsymbol{B}^{t}\|<\varepsilon$ 或者 $t>t_{\max}$，则迭代终止，否则回到②.

由此，我们得到了基于兴趣偏好的用户模糊聚类，用户类簇隶属度矩阵 $\boldsymbol{B}$ 中的元素 μ_{ij} 即表示用户 u_i 对兴趣类簇 Ψ_j 的感兴趣程度. 处在同一类簇中的用户具有相似的泛化兴趣偏好.

2.4 融合双重聚类的矩阵分解方法

1) 矩阵分解模型

矩阵分解模型是在协同过滤推荐算法中应用最为广泛的模型之一. 其主要思想是将用户-项目评分矩阵 $\boldsymbol{R}$ 近似地分解成 2 个低维矩阵乘积的形式：

$$\boldsymbol{R}\approx\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V},\quad(9)$$

其中，$\boldsymbol{U}\in\mathbb{R}^{k\times m}$，$\boldsymbol{V}\in\mathbb{R}^{k\times n}$，$k\ll\min(m,n)$. 低维矩阵 $\boldsymbol{U}$ 的第 i 列的转置 $\boldsymbol{U}_i^{\mathrm{T}}$ 与低维矩阵 $\boldsymbol{V}$ 的第 j 列 $\boldsymbol{V}_j$ 的乘积表示用户 u_i 对项目 v_j 的预测评分，因此称 $\boldsymbol{U}_i$ 为用户隐式特征向量，$\boldsymbol{V}_j$ 为项目隐式特征向量. 为了得到更优的预测评分，需要建立目标函数对预测评分矩阵和原评分矩阵的误差进行优化评估. 一般采用如下最小化平方差的目标函数[6]：

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2,\quad(10)$$

其中，I_{ij} 为指示函数，用户 u_i 对项目 v_j 产生过评分，则 $I_{ij}=1$，否则 $I_{ij}=0$. 由于用户-项目评分矩阵的稀疏程度很大，在矩阵分解过程中容易出现过拟合问题，因此需要在目标函数中添加合适的正则项.

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2+\frac{\lambda_U}{2}\|\boldsymbol{U}\|_{\text{Fro}}^2+\frac{\lambda_V}{2}\|\boldsymbol{V}\|_{\text{Fro}}^2,\quad(11)$$

其中，$\frac{\lambda_U}{2}\|\boldsymbol{U}\|_{\text{Fro}}^2$ 和 $\frac{\lambda_V}{2}\|\boldsymbol{V}\|_{\text{Fro}}^2$ 即为防止过拟合的正则项，$\|\cdot\|_{\text{Fro}}$ 为弗罗贝尼乌斯范数(Frobenius norm)，$\lambda_U(\lambda_U>0)$ 和 $\lambda_V(\lambda_V>0)$ 均为调整正则化程度的系数.

2) 融合双重聚类的矩阵分解模型

在现实生活中，我们所做的决定往往受到好友或者领域权威人士的影响. 在 2.2 节、2.3 节中我们得到了用户社交网络社区聚类信息和用户泛化兴趣偏好聚类信息，其中前者将相互影响并且特性相同的用户聚类在一起，后者将多领域兴趣偏好相似的用户聚类在一起. 显然目标用户与同一集合中的用户的相似度要高于与之不共享任一集合的用户的相似度，用户的兴趣偏好和与其同在一个集合中的所有用户的平均兴趣偏好接近，并且用户对不同集合的感兴趣程度不同. 基于以上假设，我们在文献[20]中提出的矩阵分解模型基础上进行改进，引入新的正则项：

$$\lambda_Z\sum_{i=1}^{m}\sum_{h=1}^{c_n}I_{ih}^{N}S_{ih}\sum_{g=1}^{c_p}I_{ig}^{P}\times Z_{ig}\left\|\boldsymbol{U}_i-\frac{1}{|\Omega_{h,g}(i)|}\sum_{f\in\Omega_{h,g}(i)}\boldsymbol{U}_f\right\|_{\text{Fro}}^2,\quad(12)$$

其中，$\lambda_Z(\lambda_Z>0)$ 为调整聚类正则化程度的系数. I_{ih}^{N} 为指示函数，用户 u_i 位于社交网络社区 $\gamma_h(h=1,2,\cdots,c_n)$，则 $I_{ih}^{N}=1$，否则 $I_{ih}^{N}=0$. S_{ih} 表示用户 u_i

对社交网络社区 γ_h 的感兴趣程度. I_{ig}^P 为指示函数,用户 u_i 位于兴趣偏好类簇 $\Psi_g(g=1,2,\cdots,c_p)$,则 $I_{ig}^P=1$,否则 $I_{ig}^P=0$. Z_{ig} 表示用户 u_i 对兴趣偏好类簇 Ψ_g 的感兴趣程度. $\Omega_{h,g}(i)$ 表示与用户 u_i 同在一个社交网络社区 γ_h 并且同在一个兴趣偏好类簇 Ψ_g 的用户集合.

由此,我们提出了一种新的融合双重聚类的矩阵分解模型,记为 CCMF(co-clustring based matrix factorization). 目标函数如式(13)所示.

$$L=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2+\frac{\lambda_U}{2}\|\boldsymbol{U}\|_{\mathrm{Fro}}^2+\frac{\lambda_V}{2}\|\boldsymbol{V}\|_{\mathrm{Fro}}^2+\lambda_Z\sum_{i=1}^{m}\sum_{h=1}^{c_n}I_{ih}^N S_{ih}\sum_{g=1}^{c_p}I_{ig}^P\times Z_{ig}\left\|\boldsymbol{U}_i-\frac{1}{|\Omega_{h,g}(i)|}\sum_{f\in\Omega_{h,g}(i)}\boldsymbol{U}_f\right\|_{\mathrm{Fro}}^2. \tag{13}$$

通过随机梯度下降方法得到用户隐式特征矩阵 $\boldsymbol{U}$ 和项目隐式特征矩阵 $\boldsymbol{V}$ 的局部最优解. 相应的偏导数如式(14)(15)所示.

$$\frac{\partial L}{\partial \boldsymbol{U}_i}=\sum_{j=1}^{n}I_{ij}^R\boldsymbol{V}_j(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j-R_{ij})+\lambda_U\boldsymbol{U}_i+\lambda_Z\sum_{h=1}^{c_n}I_{ih}^N S_{ih}\sum_{g=1}^{c_p}I_{ig}^P Z_{ig}\left(\boldsymbol{U}_i-\frac{1}{|\Omega_{h,g}(i)|}\sum_{f\in\Omega_{h,g}(i)}\boldsymbol{U}_f\right)+\lambda_Z\sum_{h=1}^{c_n}\sum_{g=1}^{c_p}\sum_{q\in\Omega_{h,g}(i)}\frac{S_{qh}Z_{qg}}{|\Omega_{h,g}(q)|}\times\left(\frac{1}{|\Omega_{h,g}(q)|}\sum_{f\in\Omega_{h,g}(q)}\boldsymbol{U}_f-\boldsymbol{U}_q\right), \tag{14}$$

$$\frac{\partial L}{\partial \boldsymbol{V}_j}=\sum_{i=1}^{m}I_{ij}^R\boldsymbol{U}_i(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j-R_{ij})+\lambda_V\boldsymbol{V}_j. \tag{15}$$

通过不断迭代,沿梯度下降方向更新 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 中的元素直至收敛来训练模型.

3 实验与评估

3.1 数据集

本文使用 Yelp 数据集对算法进行测试. Yelp.com 是全球最大的本地商家点评网站之一,它不但允许用户对商家进行点评或者评分,还是一个社交特征明显的互联网公司,鼓励用户之间积极互动,其平台上的用户能与其他用户建立双向确认的好友关系. 我们在本文中使用的 Yelp 数据集是 Qian 等人在文献[21]中使用的数据集. 数据集由 8 351 位用户、84 653 个项目、263 777 条项目评分信息以及 524 120 条用户双向好友关系信息组成. 其中,项目评分为[1,5]之间的整数,所有项目总共分为 8 个类别. 表 1 中是对 Yelp 数据集中分类别信息的统计数据.

Table 1 Statistic of Per Category

表 1 数据集分类别统计量

Category	User Count	Item Count	Rating Count	Sparsity
Active Life	5 328	7 492	24 385	6.109E−04
Beauty and Spas	5 466	8 494	21 344	4.597E−04
Home Services	2 500	3 213	5 182	6.451E−04
Hotels & Travel	4 712	5 883	21 658	7.813E−04
Night Life	5 074	21 267	99 878	9.256E−04
Pets	1 624	1 672	3 093	1.139E−03
Restaurants	2 000	32 725	91 946	1.405E−03
Shopping	3 000	16 154	33 352	6.882E−04

3.2 对比算法

为了验证本文提出的算法与其他算法在推荐结果准确度上的差别,我们选择 5 种算法作为对比算法进行实验.

BaseMF:文献[6]提出的适用于推荐系统的矩阵分解基本模型,没有加入用户的社交关系信息或项目类别信息.

SocialMF:文献[11]提出的融入用户信任关系信息的矩阵分解模型,假设用户向量是由其好友的用户向量决定的,在优化分解过程中引入了信任传播的概念.

SoReg:文献[20]首次提出了在矩阵分解模型中加入社交化正则项的概念,通过加入社交化正则项使得用户的偏好与其好友偏好的平均值相似.

MFC:文献[12]在矩阵分解模型中引入了重叠社区发现算法,在 SoReg 算法的基础上区分了用户所在社区不同的差异.

CircleCon:文献[13]依据针对不同的项目类别,用户对其好友的信任关系存在差异的假设,在 SocialMF 算法的基础上根据项目类别划分了用户信任网络.

3.3 评价指标

本文使用五重交叉验证作为实验方法. 将数据集随机分为 5 份,每次选择其中的 4 份即数据集的 80%作为训练集,选择余下的一份即数据集的 20%作为测试集,将 5 次的评估结果取平均值得到最终的评估数据.

由于本文提出的协同过滤算法的目标是为了提高推荐结果的准确度,因此我们采用平均绝对误差(mean absolute error, *MAE*)和均方根绝对误差(root mean square error, *RMSE*)作为实验的评估方法.

$$MAE = \frac{\sum_{(i,j)\in R_{\text{test}}} |R_{ij} - \hat{R}_{ij}|}{|R_{\text{test}}|}, \quad (16)$$

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in R_{\text{test}}} (R_{ij} - \hat{R}_{ij})^2}{|R_{\text{test}}|}}, \quad (17)$$

其中，R_{test}为测试集中所有的用户和项目的集合，R_{ij}表示用户 u_i 对项目 v_j 的真实评分，$\hat{R}_{ij}$表示用户 u_i 对项目 v_j 的预测评分. $|R_{\text{test}}|$表示测试集中的评分数量. MAE 值和 $RMSE$ 值越小表示推荐结果越准确.

3.4 实验结果及分析

1）确定兴趣类簇 l 的值

兴趣类簇 l 的值代表了根据所有用户的行为记录以及项目的类别信息划分的泛化兴趣类簇数量. 在实验中我们选择将 l 的值从 5～25 以步长 5 增加，记录推荐结果的 MAE 值和 $RMSE$ 值随不同 l 值的变化情况. 为了更加清楚地了解 l 值对推荐结果产生的影响，我们根据正则项系数 λ_Z 值的不同，分 5 组进行实验. 实验结果如图 2 所示：

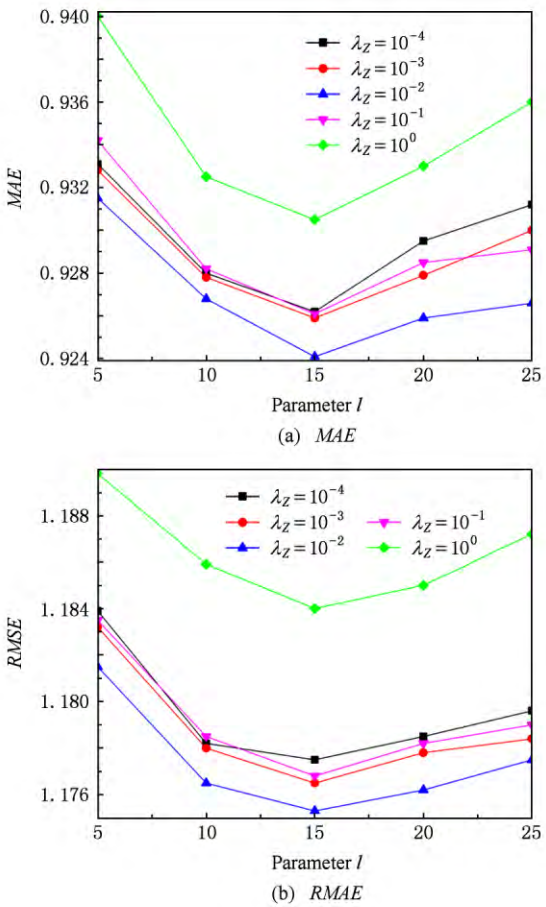

Fig. 2 Impact of parameter l.

图 2 兴趣类簇 l 的值对准确度的影响

从图 2 中可以看到，对于不同的 λ_Z 值，推荐结果的准确度随不同 l 值的变化情况基本相同，使用 l 值过大或者过小都会对推荐结果产生负面的影响. 当 $l=15$ 时，推荐结果的 MAE 值和 $RMSE$ 值同时达到最小. 分析其可能原因，如果兴趣类簇的数量设定得过小，经过模糊聚类划分出的类簇结果并不能清晰地划分用户在不同兴趣爱好的层面的分布情况；而兴趣类簇的数量设定得过大时，经过模糊聚类划分出的类簇过多，可能削弱了其对用户泛化兴趣爱好的表达.

2）确定正则项系数 λ_Z 的值

正则项系数 λ_Z 表示用户社交网络重叠社区信息以及兴趣偏好模糊聚类信息在矩阵分解模型中参与的比重，当 $\lambda_Z=0$ 时本文提出的模型即相当于基本的矩阵分解模型. 将 λ_Z 的值分别取值为{0.0001, 0.001, 0.01, 0.1, 1}进行实验. 记录推荐结果的 MAE 值和 $RMSE$ 值随不同 λ_Z 值的变化情况. 同样地，为了更加清楚地了解 λ_Z 值对推荐结果产生的影响，我们根据不同的兴趣类簇 l 的值，分 5 组进行实验. 如图 3 所示：

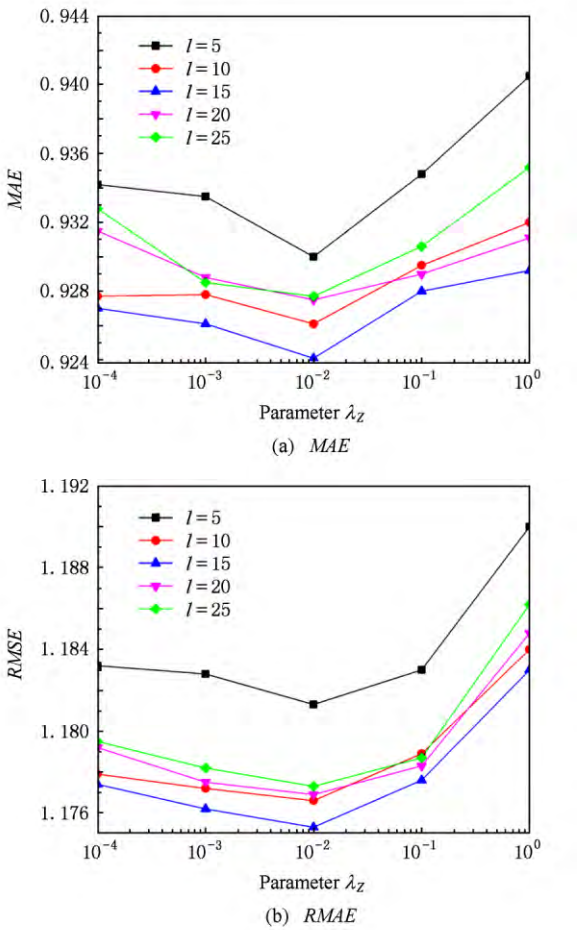

Fig. 3 Impact of parameter λ_Z.

图 3 正则项系数 λ_Z 的值对准确度的影响

从图 3 中可以看到,对于不同的 l 值,推荐结果的准确度随不同 λ_Z 值的变化情况基本相同. 当 λ_Z 取值较小时,推荐结果的 MAE 值和 $RMSE$ 值相对较高,随着 λ_Z 取更大的值,MAE 值和 $RMSE$ 值减小,当 $\lambda_Z=0.01$ 时同时达到最优的准确度. 而继续增加 λ_Z 的值后,准确度再次降低. 分析其可能原因,当 λ_Z 取值过小时,算法引入的附加信息对结果造成的影响微弱,不足以体现其在获得更优的用户隐式特征向量中所做的贡献;而 λ_Z 取值过大时,附加信息在优化向量过程中的权重过大造成过度影响.

3) 不同推荐算法的推荐效果对比

通过之前实验的对比和分析,可知当兴趣类簇 $l=15$,正则项系数 $\lambda_Z=0.01$ 时,本文提出的 CCMF 算法能够获得最优的推荐结果. 为了进一步验证 CCMF 算法的有效性,我们将 CCMF 算法与 3.2 节中介绍的相关算法进行了对比实验. 首先通过五重交叉验证确定实验中所有算法的参数. 常规正则项系数 λ_U 和 λ_V 均取值为 0.01,用户和项目的隐式特征向量维数均取 10. 在 SocialMF, SoReg, MFC, CircleCon 算法中,社交正则项系数 λ_Z 分别设为 0.01,0.01,0.001,0.01. SoReg 中的 β 值设为 0.5. 实验结果如表 2 所示:

Table 2 Comparison of CCMF and Other Methods

表 2 CCMF 算法与其他算法的实验结果对比

Method	*MAE*	*RMSE*
BaseMF	1.2095	1.6189
SocialMF	1.1039	1.4446
SoReg	0.9917	1.2726
MFC	0.9458	1.1983
CircleCon	0.9632	1.2209
CCMF	**0.9241**	**1.1753**

从表 2 中可以看到,本文提出的融合社区结构和兴趣聚类的协同过滤算法相较于其他算法,推荐结果的准确度更高. 分析其可能的原因,BaseMF 算法由于没有考虑任何用户-项目评分矩阵以外的信息所以推荐效果最差. SocialMF, SoReg, MFC 算法没有同时利用用户社交信息和项目类别信息. CircleCon 算法直接将用户根据其评分过的项目所属类别进行划分,然而单个项目类别下的用户的社交关系可能存在数据稀疏问题. 本文提出的 CCMF 算法一方面使用用户重叠社区信息,一定程度上缓解了用户的社交关系的稀疏性问题;另一方面,针对泛化兴趣偏好的模糊聚类为用户隐式特征向量提供了更有利的约束条件. 因此能够得到更好的推荐效果.

4 结束语

基于直接社交关系的传统社交化推荐算法面临用户社交信息稀疏的问题,而且没有考虑用户兴趣偏好对社交关系造成的影响,导致推荐结果不理想. 为了解决这一缺陷,本文提出了一种在矩阵分解模型中融合社区结构和兴趣聚类的协同过滤推荐算法. 利用重叠社区发现算法挖掘用户的社交关系层面的聚类,同时根据项目类别信息设计算法实现对用户兴趣偏好的模糊聚类,并且能够量化用户对不同社区或类簇的感兴趣程度. 通过实验证明该算法比现有算法能够得到更优的推荐结果.

参 考 文 献

[1] Kantor P B, Rokach L, Ricci F, et al. Recommender Systems Handbook [M]. Berlin: Springer, 2011: 1-2

[2] Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques [J/OL]. Advances in Artificial Intelligence, 2009 [2016-02-25]. http://dl.acm.org/citation.cfm?id=1722966

[3] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749

[4] Krebs V. Social network analysis: An introduction [J/OL]. 2015 [2016-02-25]. http://www.orgnet.com/sna.html

[5] Yang X, Guo Y, Liu Y, et al. A survey of collaborative filtering based social recommender systems [J]. Computer Communications, 2014, 41: 1-10

[6] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey [J]. Knowledge-Based Systems, 2013, 46: 109-132

[7] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42(8): 30-37

[8] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [C] //Proc of the 20th Neural Information Processing Systems(NIPS'07). Cambridge: MIT Press, 2007

[9] Ma H, Yang H, Lyu M R, et al. Sorec: Social recommendation using probabilistic matrix factorization [C] //Proc of the 17th ACM Conf on Information and Knowledge Management(CIKM'08). New York: ACM, 2008: 931-940

[10] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'09). New York: ACM, 2009: 203-210

[11] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks [C] //Proc of the 4th ACM Conf on Recommender Systems (RecSys'10). New York: ACM, 2010: 135-142

[12] Li H, Wu D, Tang W, et al. Overlapping community regularization for rating prediction in social recommender systems [C] //Proc of the 9th ACM Conf on Recommender Systems(RecSys'15). New York: ACM, 2015: 27-34

[13] Yang X, Steck H, Liu Y. Circle-based recommendation in online social networks [C]//Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'14). New York: ACM, 2012: 1267-1275

[14] Guo Lei, Ma Jun, Chen Zhumin. Trust strength aware social recommendation method [J]. Journal of Computer Research and Development, 2013, 50(9): 1805-1813 (in Chinese)
(郭磊, 马军, 陈竹敏. 一种信任关系强度敏感的社会化推荐算法[J]. 计算机研究与发展, 2013, 50(9): 1805-1813)

[15] Massa P, Avesani P. Trust-aware collaborative filtering for recommender systems [C] //Proc of on On the Move to Meaningful Internet Systems Confederated Int Conf. Berlin: Springer, 2004: 492-508

[16] Fortunato S. Community detection in graphs [J]. Physics Reports, 2010, 486(3): 75-174

[17] Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach [C] // Proc of the 6th ACM Int Conf on Web search and Data Mining(WSDM'13). New York: ACM, 2013: 587-596

[18] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm [J]. Computers & Geosciences, 1984, 10(2): 191-203

[19] Xu R, Wunsch D. Survey of clustering algorithms [J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678

[20] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 287-296

[21] Qian X, Feng H, Zhao G, et al. Personalized recommendation combining user interest and social circle [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26 (7): 1763-1777

Guo Hongyi, born in 1992. Master. His main research interests include recommender system and data mining (king-guo@sjtu.edu.cn).

Liu Gongshen, born in 1974, PhD, associate professor and master supervisor. Member of China Computer Federation. His research interests include information security and data mining (lgshen@sjtu.edu.cn).

Su Bo, born in 1972. Associate professor and master supervisor. His research interests include image processing and machine learning (subo@sjtu.edu.cn).

Meng Kui, born in 1973. PhD. Her research interests include information security and privacy protection (mengkui@sjtu.edu.cn).