

融合重叠社区正则化及隐式反馈的协同过滤方法

李翔锟^{1,2*}, 贾彩燕^{1,2}

(1. 北京交通大学计算机与信息技术学院, 北京 100044; 2. 交通数据分析与挖掘北京市重点实验室(北京交通大学), 北京 100044)
(* 通信作者电子邮箱 18120385@bjtu.edu.cn)

摘要: 针对目前推荐系统存在的数据稀疏和冷启动等问题, 提出了一种融合重叠社区正则化及隐式反馈的协同过滤方法(OCRIF), 该方法不仅考虑了用户在社交网络中的社区结构, 而且将用户评分信息与社交信息的隐式反馈融入推荐模型之中。此外, 由于网络表示学习可以有效学习节点在社交网络的全局结构上的近邻信息, 提出了一种网络表示学习增强的OCRIF(OCRIF+), 该方法结合社交网络中用户在网络中的低维表示与用户-商品特征, 能更有效地刻画用户之间的相似性及用户对兴趣社区的归属感。多个真实数据集上的实验结果显示: 所提出的方法的推荐效果优于同类方法, 与TrustSVD方法相比, 在FilmTrust、DouBan以及Ciao数据集上, 该方法的均方根误差(RMSE)分别下降了2.74%、2.55%以及1.83%, 平均绝对误差(MAE)分别下降了3.47%、2.97%以及2.40%。

关键词: 协同过滤; 推荐系统; 社交网络; 网络嵌入; 重叠社区

中图分类号: TP181 **文献标志码:** A

Collaborative filtering method fusing overlapping community regularization and implicit feedback

LI Xiangkun^{1,2*}, JIA Caiyan^{1,2}

(1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;
2. Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

Abstract: Aiming at the problems of data sparsity and cold start in the current recommendation system, a collaborative filtering method fusing Overlapping Community Regularization and Implicit Feedback (OCRIF) was proposed, which not only considers the community structure of users in the social network, but also integrates the implicit feedback of user rating information and social information into the recommendation model. In addition, as network representation learning can effectively learn the nodes' neighbor information on global structure of social network, a network representation learning enhanced OCRIF (OCRIF+) was proposed, which combines the low dimensional representation of users in social network with user commodity features, and can represent the similarity between the users and the membership degrees of the users to the interest communities more effectively. Experimental results on multiple real datasets show that the proposed method is superior to the similar methods on the recommendation effect. Compared with TrustSVD (Trust Support Vector Machine) method, the proposed method has the Root Mean Square Error (RMSE) decreased by 2.74%, 2.55% and 1.83% respectively, and Mean Absolute Error (MAE) decreased by 3.47%, 2.97% and 2.40% respectively on FilmTrust, DouBan and Ciao datasets.

Key words: collaborative filtering; recommendation system; social network; network embedding; overlapping community

0 引言

随着科学技术的快速发展,海量的数据充斥着人们的生活,人们很难在海量数据中提取到有用的信息。推荐系统逐渐成为人们获取个性化信息的有力工具^[1-2],并且已经成功应用到各行各业,如电影推荐(Netflix)、产品推荐(Amazon)以及音乐推荐(Last.fm)等。数据稀疏以及冷启动问题^[3]是推荐系统面临的两座大山,传统的推荐模型只考虑了用户与商品的评分矩阵^[4],但是现实中用户-商品评分矩阵具有高度稀疏而且分布不均匀的特点,仅通过分析评分矩阵进行推荐的效果

不佳。一些研究者开始在传统推荐系统中加入辅助信息来提高性能,其中最经典的是用户在社交平台上的社交信息^[5-7]。随着社交媒体的不断发展,越来越多的用户可以参与到在线活动当中,用户直接产生了大量的社交关系。依据社交关系理论^[8-9],在社交网络中拥有较强社交关系的用户之间往往具有相似的偏好,这对借助社交信息提高推荐性能提供了理论基础。在此假设下,已经有不少社交推荐模型相继被提出。实验表明:融合社交信息的推荐算法可以提高推荐模型的性能,在一定程度上缓解数据稀疏问题。尽管如此,目前常见的社交推荐方法仍存在一些问题。首先不仅用户-商品评级矩

收稿日期:2020-05-31;修回日期:2020-09-07;录用日期:2020-09-14。

基金项目:国家自然科学基金资助项目(61876016,61632004);中央高校基本科研业务费专项资金资助项目(2018JBZ006)。

作者简介:李翔锟(1996—),男,山东济宁人,硕士研究生,主要研究方向:机器学习、推荐系统;贾彩燕(1976—),女,广西北宁人,教授,博士,主要研究方向:数据挖掘、社会计算。

阵十分稀疏,用户社交关系矩阵也十分稀疏,直接利用稀疏的社交关系网络并不能很好地获取用户之间的信任程度。其次,大多数的协同过滤方法只考虑了社交关系网络中的直接邻居,忽略了用户之间的间接社交关系。最后,一些模型只是考虑了用户与商品评分矩阵和用户社交矩阵的显式反馈而没有考虑其隐式反馈信息。如何充分利用用户在社交网络上的社交信息(直接邻居和间接社交关系)并在模型中全面地考虑用户社交关系和评分矩阵中隐反馈值得关注。

针对以上问题,本文提出了一种融合重叠社区正则化及隐式反馈的协同过滤方法(collaborative filtering method fusing Overlapping Community Regularization and Implicit Feedback, OCRIF)。该方法首先利用社区发现算法将用户划分到不同社区并允许不同用户可归属于不同的社区,以充分挖掘用户潜在的共同兴趣社区;其次,OCRIF模型中融入了不同社区内的社交隐式反馈,及商品隐表示一致性约束;最后,结合网络表示学习模型^[10-13],本文给出了更有效的用户-社团隶属度计算方法以及用户之间相似度计算方法,将OCRIF模型扩展为OCRIF+。在3个公开的数据集上实验表明:本文提出的两个方法要优于以往的同类推荐算法。

1 相关工作

目前,借助社交信息来提高推荐模型性能的方法,如基于矩阵分解的方法^[14-18]、基于近邻的方法^[19]以及其他社交协同过滤方法^[20-25]相继被提出,其中基于矩阵分解的社交推荐模型因其可扩展性好、算法灵活性强等特点成为目前主流的方法。基于近邻的方法首先借助一定的相似性度量来选择与目标用户(或目标商品)相近的用户(或商品),然后根据邻居的得分来预测目标得分。矩阵分解方法依据融合社交信息的矩阵分解模型构建方式的不同,可以分为:基于矩阵分解的共享用户隐空间方法以及基于矩阵分解的融合社交约束的方法。前者是同步分解用户-商品评分矩阵和社交关系矩阵,两者分享用户特征向量;后者是使评分矩阵得到的用户特征向量受限于社交关系矩阵中的关联用户信息。

SoRec(Social Recommendation)^[14]是第一个利用矩阵分解的社交推荐方法,同时也是第一个基于矩阵分解的共享用户隐空间方法,它将社交关系矩阵考虑在内,同步分解评分矩阵和社交关系矩阵到同一个隐空间内,取得了不错的效果。Guo等^[16]在SVD++(Support Vector Machine++)^[26]的基础上提出了TrustSVD模型,这也是一种基于矩阵分解的共享用户隐空间方法,该方法不仅考虑了评分信息和社交矩阵的显式影响,还加入了隐式反馈信息。具体的,TrustSVD不仅将全局评分均值、用户对商品评分相对于全局均值的偏差以及商品评分相对于全局评分的偏差考虑进来,而且考虑了社交用户对用户特征向量的影响因子以及商品对用户特征向量的影响因子等因素。Ma等^[17]提出了SoReg(Social Regularization)模型,是一种基于矩阵分解的融合社交约束型方法,该模型假设用户虽然与其信任的好友兴趣相近,但对于不同好友应该加以区分,提出了带有社交正则化的推荐模型。鉴于已有的基于社交信息的矩阵分解模型只利用了社交网络中直接相连的邻居信息,没有利用社交网络中存在的社区结构,而同一社区的用户倾向于具有相似的兴趣,Li等^[18]提出了重叠社区正则化的推荐模型——MFC(overlapping Community regularization into Matrix Factorization),该方法利用已有的社区发现算法,

将用户划分到不同的社区中,相同社区的用户具有相似的兴趣,并允许一个用户可以被分配到不同的社区(用户存在兴趣多样性),提高了推荐模型的精度。可见,不同的协同过滤方法各有特点,适用于不同的场景,如何同时发挥不同方法的优势显得尤为重要。另外,现有的社交推荐方法在融入社交信息时,需要计算用户之间的权重约束,如用户-用户相似性权重及用户-社区相似性权重,而用户间的相似性常由用户-商品的评分矩阵信息来计算,没有充分使用社交关系矩阵中包含的信息,而现有的网络表示学习模型^[10-13]可以将网络中的节点表示成低维、实值、稠密的向量形式,考虑全局信息,使得得到的向量形式可以在向量空间中具有表示以及推理的能力,由此得到启发,本文借助该方法重新定义相似度计算方法。目前,常见的网络表示学习有DeepWalk(Deep random Walk)^[10]、LINE(Large-scale Information Network Embedding)^[11]、node2vec(node to vector)^[12]以及SDNE(Structural Deep Network Embedding)^[13]等方法;其中node2vec^[12]是一种改进随机游走策略的网络表示学习方法,该方法设置了偏置参数来控制模型更倾向于深度优先搜索还是广度优先搜索,具有很好的网络表示学习能力。

2 相关协同过滤算法

基于协同过滤的推荐方法有很多种,其中基于矩阵分解的推荐模型因其算法灵活性强、推荐性能好等特点,是当前推荐系统最主流的方法。下面详细介绍与本文密切相关的两个矩阵分解方法:TrustSVD模型^[16]和MFC^[18]模型。

假设有 m 个用户, n 个商品,令 $\mathbf{R} = \{R_{ij}\}_{m \times n}$ 表示用户对商品的评级矩阵,其中评级 R_{ij} 表示用户 u_i 对商品 v_j 的评级。此外,定义 $I_i \in I$ 为用户 u_i 评级过的商品集合,并且,假设社交网络由 $Q = \langle \mathbf{u}, \mathbf{e} \rangle$ 表示,其中 \mathbf{u} 表示 m 个用户节点集合, \mathbf{e} 表示边的集合,令 $\mathbf{G} = \{G_{ik}\}_{m \times m}$ 表示用户与用户之间的社交关系矩阵,其中 G_{ik} 表示用户 u_i 与用户 u_k 在社交关系矩阵 \mathbf{G} 中的值。

传统基于矩阵分解的方法是将评级矩阵分解为 d 维的用户特征矩阵 $\mathbf{U} \in \mathbf{R}^{d \times m}$ 和商品特征矩阵 $\mathbf{V} \in \mathbf{R}^{d \times n}$,其中 $d \ll \min(m, n)$ 。向量 \mathbf{U}_i 和 \mathbf{V}_j 分别表示用户 u_i 和商品 v_j 的 d 维潜在特征向量,其中 I_{ij}^R 为指示函数,若 $I_{ij}^R = 1$ 则表明 $R_{ij} \neq 0$ 。用户和商品的潜在特征向量可以通过最小化目标函数(式(1))来学习,并加入用户和商品特征向量的正则化项以防止过拟合。

$$J_{MF} = \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (R_{ij} - \mathbf{U}_i^T \mathbf{V}_j)^2 + \alpha_r (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \quad (1)$$

2.1 TrustSVD模型

TrustSVD模型^[16]是Guo等在SVD++^[26]的基础上得到的,它将显式的信任关系加入到隐式反馈中。其预测模型如下:

$$\hat{R}_{i,j} = b_i + b_j + \mu + \left(\mathbf{U}_i + |\mathbf{I}_i|^{-\frac{1}{2}} \sum_{p=1}^n I_{ip}^U \mathbf{y}_p + |\mathbf{T}_i|^{-\frac{1}{2}} \sum_{k=1}^m I_{ik}^Q \mathbf{W}_k \right)^T \mathbf{V}_j \quad (2)$$

其中: $\hat{R}_{i,j}$ 为 R_{ij} 的预测评分, b_i 是用户 u_i 评分的偏差, b_j 是商品 v_j 评分的偏差, μ 是全局评级的均值, \mathbf{I}_i 表示用户 u_i 评价过的商品集合, \mathbf{y}_p 表示商品 v_p 对用户特征向量的影响因子, \mathbf{T}_i 表示用户 u_i 的社交关联邻居用户集合, \mathbf{W}_k 表示用户 u_k 对用户特征向量的影响因子(社交特征向量),其中 I_{ip}^U, I_{ik}^Q 为指示函数,若

$I_{ip}^U = 1$ 则表明 $p \in I_i$, 若 $I_{ik}^Q = 1$ 则表明 $k \in T_i$ 。

对于其中的用户和商品的偏置项,它是基于这样的假设:某些用户会自带一些特质,比如天生愿意给别人好评、心慈手软或比较好说话,有的人就比较苛刻,如总是评分不超过3分(5分满分);同时也有一些这样的商品,一被生产便决定了它的地位,有的比较受人们欢迎,有的则被人嫌弃。而对于隐式反馈信息,用户对于商品的历史评分记录或者浏览记录可以从侧面反映用户的偏好,比如用户对某个商品进行了评分,可以从侧面反映他对于这类商品感兴趣。同理对于用户社交关系信息也是基于这样的假设加入到隐式反馈中。

TrustSVD^[16]模型的整体目标函数为:

$$J_{\text{TrustSVD}} = \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (R_{ij} - \hat{R}_{i,j})^2 + \alpha_u \sum_{i=1}^m \sum_{k=1}^n I_{ij}^S (G_{ik} - U_i^T \mathbf{W}_k)^2 + \alpha_r (\varphi(\mathbf{b}_i, \mathbf{b}_j, \mathbf{U}_i, \mathbf{V}_j, \mathbf{y}_p, \mathbf{W}_k)) \quad (3)$$

其中, G_{ik} 表示用户 u_i 与用户 u_k 在社交关系矩阵 \mathbf{G} 中的值, $\varphi(\mathbf{b}_i, \mathbf{b}_j, \mathbf{U}_i, \mathbf{V}_j, \mathbf{y}_p, \mathbf{W}_k)$ 表示对正则项的约束。其中 I_{ij}^S 为指示函数,若 $I_{ij}^S = 1$ 则表明 $S_{ik} \neq 0$, TrustSVD 针对不同的正则化参数采用不同的惩罚权重,评分信息更稀疏的用户和商品因为更可能过拟合而施加了更强的正则化约束。

2.2 MFC模型

除了上面介绍的利用直接社交关系的方法外,还有一些利用间接社交关系的方法,经典的如 Li 等^[18]提出的重叠社区正则化的融合社交信息推荐模型——MFC。该模型利用已有重叠社区发现算法,挖掘整个社交网络中的社区信息(获取用户-社团的隶属关系),将用户划分到不同社区中,在相同社区内的用户具有相似的偏好假设下对模型施加社交约束。

MFC^[18]模型的目标函数如下:

$$J_{\text{MFC}} = \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (R_{ij} - U_i^T \mathbf{V}_j)^2 + \alpha_r (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \alpha_u \sum_{i=1}^m \sum_{h=1}^l I_{ih}^Z Z_{ih} \sum_{w=1}^m I_{ih}^M S_{iw} \|U_i - U_w\|_F^2 \quad (4)$$

其中: l 代表社区个数, I_{ih}^Z 和 I_{ih}^M 为指示函数。若 $I_{ih}^Z = 1$ 则表明用户 u_i 属于社区 h ; 若 $I_{ih}^M = 1$ 则表明 $u_w \in \mathbf{M}_{ih}^U$, \mathbf{M}_{ih}^U 表示与用户 u_i 共同隶属于社区 h 的用户集合。 $\|U_i - U_w\|_F^2$ 保证了用户 u_i 与用户 u_w 之间的特征向量尽可能相近。 α_r 和 α_u 是自由参数,均大于0。 S_{iw} 表示用户 u_i 与用户 u_w 的皮尔逊相关系数(Pearson Correlation Coefficient, PCC)^[27], Z_{ih} 为社区所有成员的均值向量 \mathbf{C}_h 与用户 u_i 的皮尔逊相关系数,其中 S_{iw} 的 PCC 的计算公式如下:

$$S_{iw} = \frac{\sum_{k=1}^n I_{iwp}^U (R_{i,k} - \bar{R}_i)(R_{w,k} - \bar{R}_w)}{\sqrt{\sum_{k=1}^n I_{iwp}^U (R_{i,k} - \bar{R}_i)^2} \sqrt{\sum_{k=1}^n I_{iwp}^U (R_{w,k} - \bar{R}_w)^2}} \quad (5)$$

其中, $I_{iwp}^U = 1$ 代表 $k \in I_i \cap I_w$, 使用函数 $f(x) = (x+1)/2$ 将 PCC 相似性映射到区间 $[0, 1]$ 。社区均值向量 \mathbf{C}_h 为社区内所有用户向量的均值,其计算公式如下:

$$\mathbf{C}_h = \sum_{i=1}^m I_{ih}^Z u_i \quad (6)$$

而用户-社团隶属信息 Z_{ih} 的计算公式为:

$$Z_{ih} = \text{PCC}(\mathbf{C}_h, u_i) \quad (7)$$

3 融合重叠社区正则化及隐式反馈的协同过滤方法——OCRIF+

本章将提出融合重叠社区正则化的协同过滤方法 OCRIF 以及基于网络表示学习增强的 OCRIF+。

3.1 OCRIF模型

上面介绍了 TrustSVD^[16]模型,它虽然将显式的评分信息和信任关系加入到隐式反馈中,但该方法只考虑了直接相联的用户,没有考虑用户信任关系的社区传播效应。而 MFC^[18]模型,虽然它提出了一种重叠社区正则化的方法,将全局关系分割为粒度更小的社区关系,但是它没有考虑评分信息以及信任信息的隐式反馈。由此得到启发,本文提出了一种在引入重叠社区正则化的同时考虑更多维度隐式反馈信息(用户评分隐式反馈及用户社区隐式反馈)的方法,利用社区社交信息作为辅助信息,缓解了评分矩阵数据稀疏的问题,得到了如下 OCRIF 模型。

OCRIF 模型的预测评分为:

$$\hat{R}_{i,j} = b_i + b_j + \mu + \left(U_i + |I_i|^{-\frac{1}{2}} \sum_{p=1}^n I_{ip}^U \mathbf{y}_p + \sum_{h=1}^l I_{ih}^Z |M_{ih}^U|^{-\frac{1}{2}} \sum_{k=1}^m I_{ik}^Q \mathbf{W}_k \right)^T \mathbf{V}_j \quad (8)$$

其中, \mathbf{M}_{ih}^U 表示与用户 u_i 共同隶属于社区 h 的用户集合, U_w 表示在社区 h 中用户 u_w 对用户 u_i 的社区隐式影响因子(即用户 u_w 的特征向量),所以 $U_w^T \mathbf{V}_j$ 可以解释为受用户 u_i 信任的用户预测的评分,即受托人对用户 u_i 预测评分的影响。换句话说, $U_w^T \mathbf{V}_j$ 表示与用户 u_i 属于同社区的用户 u_w 对评分 R_{ij} 的影响。与评级类似,用户 u_i 的特征向量可以用所信任的同社区的用户集合来解释,即 $\sum_{h=1}^l I_{ih}^Z |M_{ih}^U|^{-\frac{1}{2}} \sum_{k=1}^m I_{ik}^Q \mathbf{W}_k$ 。与 TrustSVD^[16]模型相比,OCRIF 模型将社区的概念融入模型之中,将用户划分到不同社区中,从而在不同社区中考虑社交信息,进而将社区的社交信息加入到隐式反馈中,而不是仅使用直接相连邻居的隐式反馈信息。

本文利用现有的社区发现算法(比如 CMP (Clique Percolation Method)^[28]、BIGCLAM (Cluster Affiliation Model for Big Networks)^[29]等)挖掘社交网络中的重叠社区结构,并假设在相同社区内的用户应该具有相似的偏好,因此,类似 MFC^[18]中重叠社区正则化思想,本文将重叠社区正则化约束也加入 OCRIF 模型之中,得到如下 OCRIF 模型的目标函数:

$$L = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (R_{ij} - \hat{R}_{i,j})^2 + \frac{\alpha}{2} \sum_{p=1}^n I_{jp}^V S_{jp} \|\mathbf{V}_j - \mathbf{V}_p\|_F^2 + \frac{\beta}{2} \sum_{h=1}^l \sum_{i=1}^m I_{ih}^Z Z_{ih} \sum_{w=1}^m I_{ih}^M S_{iw} \|U_i - U_w\|_F^2 + \frac{\gamma}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \sum_i \|b_i\|_F^2 + \sum_j \|b_j\|_F^2 + \sum_p \|\mathbf{y}_p\|_F^2 \right) \quad (9)$$

其中, $\alpha, \beta, \gamma > 0$ 为自由参数。 S_{jp} 与 S_{iw} 分别代表商品 v_j 与 v_p 的皮尔逊相关系数以及用户 u_i 与 u_w 的皮尔逊相关系数,其中 I_{jp}^V 为指示函数,若 $I_{jp}^V = 1$ 则表明 $p \in N_j$, N_j 代表商品 v_j 的近邻集合。

通过对变量 $U_i, \mathbf{V}_j, b_i, b_j, \mathbf{y}_p$ 进行梯度下降,可以得到目标函数(式(9))的局部极小值。

$$\begin{cases}
\frac{\partial L}{\partial b_i} = \sum_{j=1}^n I_{ij}^R (R_{ij} - \hat{R}_{i,j}) \\
\frac{\partial L}{\partial b_j} = \sum_{i=1}^m I_{ij}^R (R_{ij} - \hat{R}_{i,j}) \\
\frac{\partial L}{\partial U_i} = \sum_{j=1}^n I_{ij}^R (R_{ij} - \hat{R}_{i,j}) \left(\sum_{p=1}^m I_{ip}^D \sum_{h=1}^l I_{ih}^F |M_{ih}^U|^{\frac{1}{2}} \right) V_j + \\
\quad \gamma U_i + \beta \sum_{i=1}^m \sum_{h=1}^l I_{ih}^Z Z_{ih} \sum_{w=1}^m I_{ih}^M S_{iw} (U_i - U_w) - \\
\quad \beta \sum_{p=1}^m I_{ip}^D \sum_{h=1}^l I_{ih}^F Z_{ph} S_{pi} (U_p - U_i) \\
\frac{\partial L}{\partial V_j} = \sum_{i=1}^m I_{ij}^R (R_{ij} - \hat{R}_{i,j}) U_i + \alpha \sum_{j=1}^n I_{ij}^R S_{jn} (R_{ij} - \hat{R}_{i,j}) + \\
\quad \gamma V_j \\
\frac{\partial L}{\partial y_p} = \sum_{j=1}^n I_{ij}^R S_{jn} (R_{ij} - \hat{R}_{i,j}) |I_i|^{\frac{1}{2}} V_j + \gamma y_p; \quad \forall p \in I_i
\end{cases} \quad (10)$$

其中, I_{ip}^D 和 I_{ih}^F 为指示函数, 若 $I_{ip}^D = 1$ 则表明 $u_p \in D$, 若 $I_{ih}^F = 1$ 则表明 $C_h \in F$, E_j 表示评论过商品 v_j 的用户集合, 且 $D = \{u_p | \exists h, u_i \in M_{ih}^U \& I_{ph}^Z = 1\}$, $F = \{C_h \in I | u_i \in M_{ph}^U \& I_{ph}^Z = 1\}$ 。

与前人提出的模型类似, 上述方法在计算用户与社团的隶属度的时候只考虑了评分信息, 没有充分地利用社交信息。进一步, 本文将在 3.2 节中将社交信息加入到用户-社团隶属度以及用户之间的相似度的计算当中, 以更好地融合用户社交信息和用户-商品评分信息。

3.2 OCRIF+模型

在 OCRIF 模型中需要计算用户 u_i 与用户 u_w 之间的相似度 S_{iw} 以及用户-社团隶属信息 Z_{ih} 的大小, 而这两组参数值对模型效果影响较大。为了更有效地利用用户社交关系信息, 本文借助网络表示学习方法, 将用户在网络中全局拓扑关系映射到一个低维、紧密的向量空间, 以更好地融合用户-商品兴趣特征信息, 给出 S_{iw} 以及 Z_{ih} 更有效地估计。

本文依据用户社交网络信息, 借助 node2vec^[12] 模型学习到用户 u_i 的低维表示 $g_i \in \mathbf{R}^d$ 以及用户 u_w 的低维表示 $g_w \in \mathbf{R}^d$ 。给定 g_i, g_w 可以计算用户-社团隶属度关系以及用户 u_i, u_w 之间的相似度。现有的网络表示学习方法已经取得了相当不错的效果, 借助网络表示学习技术捕获全局拓扑关系, 便于不同类型信息的融合, 本文选用 node2vec^[12] 模型挖掘拓扑关系, 其他网络表示学习技术也同样适用, 如 DeepWalk^[10] 等方法。

表 1 实验数据集统计

Tab. 1 Statistics of experimental datasets

数据集	#user(n)	#item(m)	#ratings	RDensity/%	\bar{m}	\bar{n}	relation	SDensity/%	\bar{s}
FilmTrust	1 641	2 071	35 487	1.044	23	17	1 853	0.069 0	2
Douban	129 490	58 541	16 830 839	0.010	14	5	1 692 952	0.020 1	10
Ciao	7 375	106 997	284 086	0.036	39	3	111 781	0.205 5	15

4.2 实验评价指标

为了衡量本文提出方法的推荐准确度, 本文选取了两个最常用的评价指标: 均方根误差 (Root Mean Squared Error, RMSE) 以及平均绝对误差 (Mean Absolute Error, MAE)。方法虽然不同, 但是这两种指标都是通过计算预测值与真实值之间的误差来衡量推荐精度的, 值越小推荐精度越高。具体计算方法如下:

$$MAE = \frac{\sum_{ij} |R_{ij} - \hat{R}_{i,j}|}{|R|} \quad (13)$$

其中, 用户之间的相似度 S_{iw} 计算公式如下:

$$S_{iw} = \lambda_s PCC(u_i, u_w) + (1 - \lambda_s) PCC(g_i, g_w) \quad (11)$$

用户-社团隶属信息 Z_{ih} 的计算公式如下:

$$Z_{ih} = \lambda_z PCC(u_i, C_h) + (1 - \lambda_z) PCC(g_i, C_h) \quad (12)$$

其中 λ_s 以及 λ_z 为超参数来平衡两种信息所占比例。

利用网络表示学习方法挖掘网络拓扑关系有望充分考虑评分信息以及社交信息, 提高相似度的计算精确度, 缓解了数据稀疏带来的影响。

3.3 算法复杂度分析

设 d 为潜在空间的维度, m 为用户数, n 为商品数, \bar{m} 为每个用户给出的平均评分数, \bar{n} 为每个用户所属社区的平均数, \bar{w} 为每个社区的平均成员数。评估目标函数的复杂度为 $O(\bar{m}nd + \bar{n}wd)$, 而计算梯度的成本是 $O(\bar{m}nd + \bar{n}wd)$ 。这些成本与 \bar{m} 和 \bar{n} 呈线性关系。由于评级矩阵非常稀疏, \bar{m} 相对较小。根据对包括信息网络 (Wikipedia)、内容共享网络 (Flickr) 和社交网络 (Facebook、Google+ 和 Twitter) 在内的真实网络的分析, \bar{n} 和 \bar{w} 也很小。

4 实验分析

4.1 数据集

本文选取了 FilmTrust、DouBan 以及 Ciao 三个数据集进行实验。这些数据集是在社交网站上运行爬虫程序获取的。具体的: FilmTrust 数据集 (<http://www.librec.net/datasets.html>) 是在电影推荐网站 FilmTrust (www.trust.mindswap.org/) 爬取获得, 用户可以根据自己的偏好对电影进行评级, 同时用户之间可以建立信任关系; DouBan 数据集 (<https://www.cse.cuhk.edu.hk/irwin.king.new/pub/data/douban>) 是在社交网站 DouBan (www.douban.com) 上抓取获得的, 包含了用户对于书籍、音乐、电影的评分信息, 同样该社交网站也允许用户添加信任朋友; Ciao 数据集 (<http://www.jiliang.xyz/trust.html>) 是在商品评价网站 Ciao (www.ciao.co.uk) 上抓取的, 允许用户对商品进行评价。不同数据集的评分范围不尽相同, FilmTrust 评分范围为 0.5~4, 步长为 0.5; Douban 以及 Ciao 评分范围为 1~5, 步长为 1。数据集统计特征如表 1 所示, 表中 #user(n) 表示数据集中的用户数, #item(m) 表示数据集中的商品数, #ratings 表示数据集中的评分数, relation 表示社交矩阵中用户之间的关系数, RDensity 表示评分矩阵中数据的稀疏程度, SDensity 表示在社交关系矩阵中数据的稀疏程度, \bar{n} 代表单个商品的平均评分用户数, \bar{m} 代表单个用户的平均评分商品数, \bar{s} 代表单个用户的平均关系数。

$$RMSE = \sqrt{\frac{\sum_{ij} (R_{ij} - \hat{R}_{i,j})^2}{|R|}} \quad (14)$$

其中: R_{ij} 代表真实评级, $\hat{R}_{i,j}$ 代表预测评级, $|R|$ 代表评级的个数。

4.3 实验对比方法及参数设置

为了验证 OCRIF 及 OCRIF+ 方法的有效性, 本文选择了 7 个在推荐系统领域比较经典的相关协同过滤方法作为对比。

1) PMF (Probabilistic Matrix Factorization)^[30]: 概率矩阵分

解模型。该模型从概率的角度对矩阵分解进行解释,是一种潜因子模型,在同一个空间内对用户和商品进行分解,仅使用了评分矩阵信息。

2)SVD++^[26]:一种基于矩阵分解的潜因子模型。该模型不仅将用户和商品的偏差考虑在内,而且考虑了用户评分信息的隐式反馈,也仅使用了评分矩阵信息。

3)TrustMF(Trust Matrix Factorization)^[15]:一种社交推荐模型。该模型假设信任关系是有向的,将社交矩阵分解为信任者特征向量以及被信任者特征向量。

4)SoReg^[17]:一种社会正则化模型。该模型通过用户相似度来控制两个用户的相近程度。

5)MFC^[18]:一个利用重叠社区正则化的社会推荐模型。该模型基于用户社交关系信息利用社区挖掘算法将用户划分到重叠社区内,并将社区正则项融入矩阵分解模型。

6)MFC+:本文提出的OCRIF+模型对于OCRIF模型的改进在于借助node2vec模型学习到用户向量的低维表示,进而更好地估计用户之间的相似度以及用户-社团隶属信息,本文将这种思想用于MFC模型,得到MFC+模型作为对比实验,验证这种思想的有效性。

7)TrustSVD^[16]:SVD++^[26]方法进行扩展模型。该模型同时考虑用户评级信息和社交信息,并且将社会隐式反馈考虑在内模型约束中。

表2对比方法中的参数均使用前人文章中所描述的最优参数设置或通过实验选择来获得最佳的结果。在本文提出的方法中,参数通过五折交叉验证进行调整。另外,根据实验结果分析,本文将用户和商品的潜在表示维数设置为10,网络表示学习中用户顶点分布式表示维数为30。在FilmTrust数据集以及Ciao数据集中 $\lambda_1=0.8, \lambda_2=0.6$;在DouBan数据集中, $\lambda_1=0.6, \lambda_2=0.5$ 。在3个数据集中,统一设置参数 $\alpha=0.0005, \beta=0.002, \gamma=0.02$ 。本文采用五折交叉验证的方式,将数据集等分为5个子数据集,每次实验,4份子数据集作为训练集,剩下的1份子数据集作为测试集,5次实验后能能保证所有子数据集都被测试,5次实验后结果取平均作为最终结果。

4.4 实验结果及实验分析

为了验证模型的有效性,本文从整体用户以及冷启动用户两个角度对实验结果进行分析。一般情况下,对于评分数量小于5的用户称为冷启动用户^[31],表2中的实验(a)与(b)分别展示了整体用户以及冷启动用户预测评分与实际评分之间的MAE以及RMSE。

表2 实验结果对比

Tab. 2 Comparison of experimental results

实验名称	数据集	FilmTrust		DouBan		Ciao	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
(a)整体用户上实验结果对比	PMF	0.87278	0.66901	0.78692	0.66470	1.01930	0.78034
	SVD++	0.85204	0.65660	0.83715	0.65267	1.02908	0.77308
	TrustMF	0.85813	0.64802	0.81927	0.64092	1.00799	0.76280
	SoReg	0.85387	0.64690	0.81014	0.63982	0.99872	0.75620
	MFC	0.84868	0.65265	0.80672	0.63028	0.99027	0.75392
	MFC+	0.83157	0.64510	0.79389	0.62719	0.98362	0.74925
	TrustSVD	0.83990	0.64969	0.79293	0.62641	1.02763	0.75929
	OCRIF	0.82601	0.63829	0.78120	0.61491	0.97815	0.74011
	OCRIF+	0.81690	0.62712	0.77271	0.60782	0.97106	0.73229
(b)冷启动用户上实验结果对比	PMF	0.93398	0.74208	1.04094	0.88506	1.07621	0.88912
	SVD++	0.91375	0.73891	0.98671	0.78637	1.09362	0.87160
	TrustMF	0.93708	0.67336	0.97325	0.77219	1.06231	0.84712
	SoReg	1.01882	0.74581	1.02537	0.81879	1.06163	0.87762
	MFC	0.90012	0.68971	0.96368	0.75972	1.03719	0.81237
	MFC+	0.87291	0.67512	0.93019	0.75003	1.02196	0.80382
	TrustSVD	0.87941	0.67629	0.93971	0.74529	1.05572	0.80612
	OCRIF	0.86267	0.65670	0.91275	0.72901	1.00569	0.79421
	OCRIF+	0.85190	0.63921	0.90021	0.71754	0.99679	0.78129

通过表2中的实验(a)对比本文提出的两个模型与其他模型的实验结果可以得到以下两个结论:

1)大多数情况下,融合社交协同过滤方法方法优于仅考虑评分信息的传统协同过滤方法。此外,本文提出的OCRIF方法明显优于PMF、TrustSVD^[16]以及MFC^[18]方法,这样证明了本文提出的将重叠社区正则化与隐式反馈信息结合模型能够提高推荐的精度。

2)通过与其他7种协同过滤方法比较,本文提出的OCRIF+方法在3个数据集上都取得了最优的效果。此外通过比较MFC与MFC+方法以及OCRIF与OCRIF+方法,后者的实验结果优于前者,这说明通过网络表示学习来改进模型的思想可以更好地利用社交信息,在一定程度上缓解数据稀疏现象,提高推荐算法的精度而且适用于其他同类模型。

正如前面所提到的,冷启动问题是推荐系统面临的主要挑战之一。为了有效评估模型解决冷启动问题的能力,本文专门抽取了冷启动用户进行实验分析。在本文选取的FilmTrust、DouBan以及Ciao三个数据集中,冷启动用户数分别为315、18943和171,分别占总数的19.2%、14.6%和2.3%。由表2中的实验(b)的实验结果可以看出,本文方法在冷启动用户上取得了较好的结果,一定程度上缓解了冷启动问题。

4.5 实验参数分析

在本文提出的两种模型中引入了几个超参数(α, β, γ 以及 λ_1, λ_2)的取值将会影响模型的效果。其中, α 和 β 分别代表商品特征矩阵相似性以及重叠社区正则化在模型中所占的比

例, λ_s 和 λ_z 分别代表用户商品矩阵信息和社交关系矩阵信息所占比例在用户之间的相似度 S_{in} 以及用户-社团隶属信息 Z_{in} 中所占比例。本文以 FilmTrust 数据集为例, 研究上述参数对模型的影响。图 1~4 分别代表 α 、 β 以及 λ_s 和 λ_z 的取值影响模型 MAE 以及 RMSE 的情况, γ 为这类模型中的常规正则项, 这里用网络搜索取最优值。由图 1 可知, 当 $\alpha=0.0005$ 时效果最好; 当 α 值过高时, 将会引入噪声削减了评分以及社会信息的比重。由图 2 可知, 当 $\beta=0.002$ 时效果最佳; 当 β 值较小时, 社会信息所占比例较小, 推荐性能降低; 但当 β 值较大时, 过多的社交信息将会引入噪声, 影响模型性能。由图 3 以及图 4 可知, 当 $\lambda_s=0.8$, $\lambda_z=0.6$ 时 MAE 以及 RMSE 最低, 参数大于 0.5 说明在计算相似度时评分矩阵使所占比例较高, 极端情况下当 $\lambda_s=\lambda_z=1$ 时, OCRIF+模型将变成 OCRIF 模型, 效果有所降低。

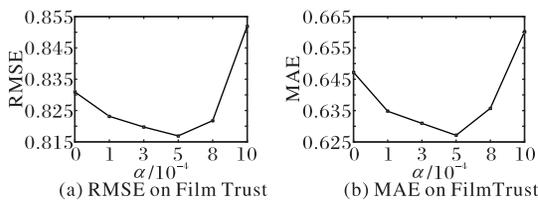


图 1 α 对 OCRIF+模型精度的影响

Fig. 1 Influence of α on OCRIF+ model accuracy

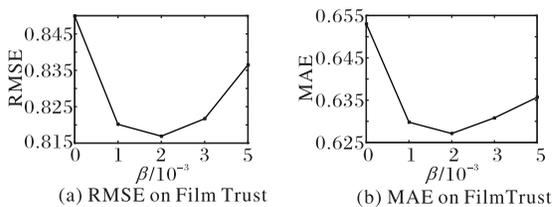


图 2 β 对 OCRIF+模型精度的影响

Fig. 2 Influence of β on OCRIF+ model accuracy

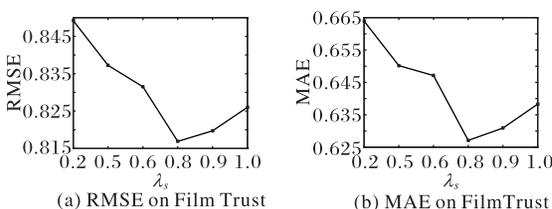


图 3 λ_s 对 OCRIF+模型精度的影响

Fig. 3 Influence of λ_s on OCRIF+ model accuracy

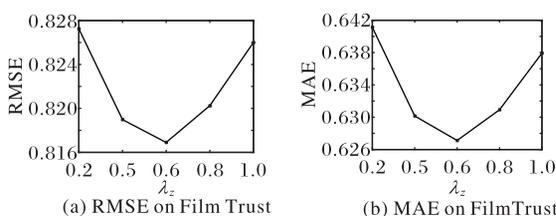


图 4 λ_z 对 OCRIF+模型精度的影响

Fig. 4 Influence of λ_z on OCRIF+ model accuracy

此外为了验证模型的合理性, 本文还单独双比了 OCRIF+模型中了各约束项的有效性的(实验结果如表 3)。表 3 中展示了 4 组实验, 非 0 的参数取之前提到的模型最优设置。由表 3 可知: 当 $\alpha=0$ 时, 在模型中去除商品的平滑项, 模型效果有所下降, 但优于 MFC^[18] 模型以及 TrustSVD^[16] 模型, 这表明在社区中加入隐式反馈信息的方法是有效的; 当 $\beta=0$ 时, 模型去除

了重叠社区正则化的部分, 大大降低了社交信息在模型中的比例, 但保留了社交的隐式反馈部分, 所以模型推荐效果还是优于 SVD++^[26] 模型的; 当 $\alpha=0$ 且 $\beta=0$ 时, 效果大大降低, 但还是保留了隐式反馈信息, 结果与 SVD++ 基本持平。通过分析发现, 本文提出的融合社区正则化及隐式反馈的协同过滤方法是合理且有效的。

表 3 OCRIF+参数有效性实验结果

Tab. 3 OCRIF+ parameter validity experiment results

参数值	FilmTrust		DouBan		Ciao	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
$\alpha=0,$ $\beta \neq 0$	0.83129	0.64672	0.78917	0.62035	0.98274	0.74792
$\alpha \neq 0,$ $\beta=0$	0.84991	0.65305	0.83284	0.64715	1.01493	0.76821
$\alpha=0,$ $\beta=0$	0.85112	0.65629	0.83592	0.65028	1.02157	0.77012
$\alpha \neq 0,$ $\beta \neq 0$	0.81690	0.62712	0.77271	0.60782	0.97106	0.73229

5 结语

本文提出了一种协同过滤方法的通用框架 OCRIF, 该框架为了区分用户之间的社交关系, 借助了重叠社区正则化的思想, 将用户划分到不同的重叠社区之中, 并且将用户评分信息与社区社交信息加入到隐式反馈当中。此外, 为了更好地挖掘用户的兴趣与社交关系, 充分利用评分矩阵以及社交网络信息, 缓解数据稀疏问题, 本文借助网络表示学习方法得到用户的分布式表示, 进而挖掘用户之间的联系, 本文将该方法称为 OCRIF+。实验结果证明了本文提出的方法优于现有的同类方法。

参考文献 (References)

- [1] RESNICK P, IACOVOU N, SUCHAK M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]// Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. New York: ACM, 1994: 175-186.
- [2] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.
- [3] JAMALI M, ESTER M. A matrix factorization technique with trust propagation for recommendation in social networks [C]// Proceedings of the 4th ACM Conference on Recommender Systems. New York: ACM, 2010: 135-142.
- [4] ANANDHAN A, SHUIB L, ISMAIL M A, et al. Social media recommender systems: review and open research issues [J]. IEEE Access, 2018, 6: 15608-15628.
- [5] TERVEEN L, MCDONALD D W. Social matching: a framework and research agenda [J]. ACM Transactions on Computer-Human Interaction, 2005, 12(3): 401-434.
- [6] TANG J, HU X, LIU H. Social recommendation: a review [J]. Social Network Analysis and Mining, 2013, 3(4): 1113-1133.
- [7] MASSA P, AVESANI P. Trust-aware recommender systems [C]// Proceedings of the 2007 ACM Conference on Recommender Systems. New York: ACM, 2007: 17-24.
- [8] MARSDEN P V, FRIEDKIN N E. Network studies of social influence [J]. Sociological Methods and Research, 1993, 22(1):

- 127-151.
- [9] SCOTT J. Social network analysis: a handbook [J]. Contemporary Sociology, 1993, 22(1): 157-169.
- [10] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710.
- [11] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding [C]// Proceedings of the 24th International Conference on World Wide Web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2015: 1067-1077.
- [12] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864.
- [13] WANG D, CUI P, ZHU W. Structural deep network embedding [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1225-1234.
- [14] MA H, YANG H, LYU M R, et al. SoRec: social recommendation using probabilistic matrix factorization [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008:931-940.
- [15] YANG B, LEI Y, LIU J, et al. Social collaborative filtering by trust [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1633-1647.
- [16] GUO G, ZHANG J, YORKE-SMITH N. TrustSVD: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 123-129.
- [17] MA H, ZHOU D, LIU C, et al. Recommender systems with social regularization [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011:287-296.
- [18] LI H, WU D, TANG W, et al. Overlapping community regularization for rating prediction in social recommender systems [C]// Proceedings of the 9th ACM Conference on Recommender Systems. New York: ACM, 2015: 27-34.
- [19] BEN KHARRAT F, ELKHLIFI A, FAIZ R. Empirical study of social collaborative filtering algorithm [C]// Proceedings of the 2016 Asian Conference on Intelligent Information and Database Systems, LNCS 9622. Berlin: Springer, 2016: 85-95.
- [20] JAMALI M, ESTER M. *TrustWalker*: a random walk model for combining trust-based and item-based recommendation [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 397-406.
- [21] JAMALI M, ESTER M. Using a trust network to improve top-N recommendation [C]// Proceedings of the 3rd ACM Conference on Recommender Systems. New York: ACM, 2009: 181-188.
- [22] MAN T, SHEN H, JIN X, et al. Cross-domain recommendation: an embedding and mapping approach [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 2464-2470.
- [23] FAN W, LI Q, CHENG M. Deep modeling of social relations for recommendation [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 8075-8076.
- [24] WU L, SUN P, FU Y, et al. A neural influence diffusion model for social recommendation [C]// Proceedings of the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2019: 235-244.
- [25] SEDHAIN S, MENON A K, SANNER S, et al. Low-rank linear cold-start recommendation from social data [C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 1502-1508.
- [26] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 426-434.
- [27] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering [C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. New York: ACM, 1998: 43-52.
- [28] PALLA G, DERÉNYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043):814-818.
- [29] YANG J, LESKOVEC J. Overlapping community detection at scale: a nonnegative matrix factorization approach [C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM, 2013: 587-596.
- [30] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization [C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2007: 1257-1264.
- [31] MASSA P, AVESANI P. Trust-aware recommender systems [C]// Proceedings of the 2007 Conference on Recommender Systems. New York: ACM, 2007: 17-24.

This work is partially supported by the National Natural Science Foundation of China (61876016, 61632004), the Fundamental Research Funds for the Central Universities (2018JBZ006).

LI Xiangkun, born in 1996, M. S. candidate. His research interests machine learning, recommendation system.

JIA Caiyan, born in 1976, Ph. D., professor. Her research interests include data mining, social computing.