ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 483－493
DOI:10. 16337/j. 1004-9037. 2020. 03. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于图嵌入模型的协同过滤推荐算法

高海燕[1]，毛　林[1]，窦凯奇[1]，倪文晔[1]，赵卫滨[1,2]，余永红[1]

(1. 南京邮电大学通达学院，扬州，225127;2. 计算机软件新技术国家重点实验室(南京大学)，南京，210023)

摘　要：传统协同过滤算法存在严重的数据稀疏和冷启动问题。利用社交网络中的丰富信息为解决传统协同过滤算法的数据稀疏和冷启动带来了契机。然而，传统基于社交网络的协同过滤算法仅利用粗粒度、稀疏的用户信任关系来改进传统协同过滤算法，即用0或1表示用户之间信任程度。另外，传统基于社交网络推荐算法仅仅集成用户之间显式信任关系，而忽略用户之间隐式的信任关系。本文提出一种基于图嵌入模型的协同过滤推荐算法，即利用图嵌入模型技术学习社交网络中用户的低维特征表示，并根据用户的低维特征表示推导用户之间细粒度的信任关系。最后，根据信任用户和相似用户对目标物品的评分权重预测用户对目标物品的评分。在真实数据集上的实验结果表明，基于图嵌入模型的协同过滤算法的性能优于传统的协同过滤算法。

关键词：协同过滤;图嵌入技术;社交网络

中图分类号：TP181　　**文献标志码**：A

Graph Embedding Model Based Collaborative Filtering Algorithm

GAO Haiyan[1], MAO Lin[1], DOU Kaiqi[1], NI Wenye[1], ZHAO Weibin[1,2], YU Yonghong[1]

(1. Tongda College, Nanjing University of Posts and Telecommunications, Yangzhou, 225127, China;2. State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing, 210023, China)

Abstract: Traditional collaborative filtering algorithms suffer from data sparsity and cold start problems. Taking advantage of rich information in social networks brings an opportunity to alleviate the problems of data sparsity and cold start. However, the traditional social network-based collaborative filtering algorithm only use the coarse-grained and sparse trust relationships to improve recommendation quality, i.e. they only utilize 0 or 1 to denote the trust relationships between users. In addition, the traditional social network based recommendation algorithms only integrate explicit trust relationships, and ignore implicit trust relationships. In this paper, we propose a graph embedding model based collaborative filtering algorithm. Specifically, we adopt the graph embedding technique to learn the low-dimensional embedded representations of users in social networks, and infer the fine-grained trust relationship between users based on the low-dimensional embedded representations. Finally, the user's rating of the target item is predicted based on the scoring weights of the target item by the trusted user and the similar one. Experimental results on the actual data sets prove that the performance of the collaborative filtering algorithm based on the graph embedding model is better than that of the traditional collaborative filtering algorithms.

基金项目：江苏省青蓝工程资助项目;江苏省高校自然科学研究(17KJB520028)资助项目;南京邮电大学校级科研基金(NY217114)资助项目;南京邮电大学通达学院科研基金(XK203XZ18002)资助项目。
收稿日期：2019-10-25;**修订日期**：2019-11-26

Key words: collaborative filtering; graph embedding model; social networks

引　言

在大数据时代,推荐系统[1]成为电子商务、社交媒体和在线新闻等应用平台中必不可少的智能组件。推荐系统根据用户历史数据挖掘用户隐式偏好,从而为用户提供个性化的服务,有效缓解了信息过载问题,成为业界的研究热点。协同过滤(Collaborative filtering,CF)算法[2-3]是推荐系统中应用最为广泛的方法。协同过滤算法通过分析用户的历史行为数据来预测用户未来的偏好。但是,数据稀疏和冷启动是限制协同过滤推荐算法性能的主要问题。

社交网络的出现,促使越来越多的推荐算法利用社交网络提供的丰富信息解决传统推荐算法的数据稀疏和冷启动等问题。SoRec[4],RSTE[5],SocialMF[6],TrustMF[7]等是典型的基于社交网络的推荐算法。然而,在原始社交网络中,信任关系往往是二值的,即仅使用0/1表示用户之间的信任关系,1表示信任,并且信任强度为1;0表示用户之间不存在信任关系。直观地,由于用户对不同人的信任强度存在差异,使得这种信任关系表示方法粒度粗。另外,原始的社交网络中,稀疏的信任关系使得可以利用的社交关系数据较少。实际上,很多用户虽然没有直接建立信任关系,但是他们具有很多共同的邻居,他们之间很可能具有较高的信任程度。在进行推荐模型建模时,考虑这种间接的、隐式的信任关系可以有效地提高推荐模型的性能。但是,传统的基于社交网络推荐算法往往忽视这种用户之间隐式的信任关系。另外,为了建模信息网络的全局结构和局部结构,研究人员提出一系列的图嵌入(Graph embedding,GE)模型[8-11]。图嵌入模型将大规模的信息网络嵌入到低维空间中,每个网络节点用低维特征向量表示。这种低维特征向量有效保存了信息网络的结构信息,可以应用于基于图的各种机器学习任务中,如节点分类、聚类、链路预测和可视化等。典型的图嵌入模型包含DeepWalk[9],Graph Factorization[10],LINE[11]等。例如,LINE[11]模型通过同时保留大规模信息网络的全局结构和局部结构的目标函数学习大规模信息网络的嵌入式表示。另外,LINE模型使用边缘采样策略,克服了经典随机梯度下降算法的局限性,适用于任意类型的网络(有向或无向,加权或非加权),并且可以很容易扩展到数百万个节点。另外,通过LINE模型学习到的节点特征表示同时保留节点之间的一阶相似度和二阶相似度。直观地,可以通过图嵌入技术学习社交网络中用户节点的低维特征表示,并集成到传统的基于社交网络推荐算法中,解决传统基于社交网络推荐算法中用户信任关系粒度粗、仅考虑显式信任关系等问题。但是,很少研究工作考虑在传统基于社交网络推荐算法中融合图嵌入技术,以提升传统基于社交网络推荐算法的性能。

针对传统基于社交网络推荐算法存在的问题,本文提出一种基于图嵌入模型的协同过滤推荐算法。首先,利用图嵌入模型技术学习社交网络中用户的低维特征表示。这种用户低维特征表示同时保存社交网络的全局结构和局部结构信息,即同时保存了用户之间的一阶信任度和二阶信任度。然后,根据用户的低维特征表示推导用户之间细粒度的信任关系。最后,以信任用户和相似用户对目标物品的评分权重平均预测用户对目标物品的评分。在真实数据集中的实验结果表明,基于图嵌入模型的协同过滤算法的性能优于传统的协同过滤算法。

1　相关研究工作

1.1　协同过滤推荐算法

目前应用最广泛的推荐技术是协同过滤推荐算法。它从用户的历史活动记录挖掘用户隐藏偏好,并根据用户隐藏偏好进行推荐。协同过滤算法主要分为3类:基于内存的协同过滤推荐算法、基于模型

的协同过滤算法和混合推荐算法[1]。

由于基于矩阵分解的推荐算法[12-13]在处理大规模数据方面有着较高的可扩展性和预测能力,因此受到业内人士的广泛关注。基于矩阵分解的推荐算法将用户和项目同时映射到两个低维的隐式因子空间中,利用用户和项目的低维特征向量的内积来预测用户对项目的评分。PMF[13],SVD++[14],NMF[15],MMMF[16]和NPCA[17]等均为典型的基于矩阵分解的推荐算法。在传统协同过滤算法中,用户-项目评分矩阵极度稀疏性严重影响了协同过滤算法的性能。例如,在用户的评分信息较少或缺失时,基于用户的协同过滤推荐算法[2]很难根据用户的评分信息寻找到与之兴趣相似的其他用户。本文利用图嵌入模型对传统的基于用户的协同过滤算法进行改进,尤其是在基于用户协同过滤算法的基础上进一步融合细粒度的用户信任关系,同时考虑用户之间的显式信任和隐式信任,减轻传统协同过滤算法中的数据稀疏和冷启动问题。

1.2 图嵌入模型

图嵌入技术将大规模信息网络嵌入到低维向量空间中,信息网络中每个节点用一个低维特征向量表示。这种低维特征向量表示保存了信息网络的结构信息。网络嵌入技术在节点分类,聚类,链路预测,可视化等机器学习任务中具有广泛的应用前景。典型的图嵌入技术包含DeepWalk[9],Graph Factorization[10]和LINE[11]等。

Graph Factorization[10]通过矩阵分解模型来学习大规模网络的嵌入式表示。但是,在Graph Factorization中使用的矩阵分解的目标函数并不是针对网络设计的,因此不能保存信息网络的全局结构信息。而且,Graph Factorization模型仅适用于无向信息网络。DeepWalk[9]通过随机游走算法来学习信息网络的嵌入式表示。然而,DeepWalk模型并没有清晰地描述保存信息网络的何种属性。另外,DeepWalk模型期望具有类似邻居结构的节点具有类似的低维特征表示,而没有考虑具有直接链接关系节点之间的低维特征向量之间的关系。DeepWalk模型仅适用于无权重网络。LINE[11]通过保留大规模信息网络的局部和全局结构的目标函数学习大规模信息网络的嵌入式表示。另外,LINE使用边缘采样策略处理经典随机梯度下降算法中存在的局限性。LINE模型适用于大规模同构信息网络,包括无向/有向、权重和无权重信息网络。另外,通过LINE模型学习到的节点特征表示同时保留节点之间的一阶相似度和二阶相似度。

鉴于LINE模型的高效性,本文采用LINE模型学习社交网络中用户节点的嵌入式表示,并将社交网络中用户节点的嵌入式表示集成到传统的协同过滤推荐算法中。

2 预备知识

2.1 推荐问题的形式化描述

协同过滤算法主要利用用户-项目评分矩阵来预测用户对项目的评分。用户-项目评分矩阵$\boldsymbol{R}_{M\times N}$包含$M$个用户集合$U=\{u_1,u_2,\cdots,u_M\}$和$N$个项目集合$I=\{i_1,i_2,\cdots,i_N\}$。$\boldsymbol{R}_{M\times N}$中的每项$r_{ui}$是用户$u$对项目$i$的评分。评分数据$r_{ui}$可取任意的实数,但通常情况下,评分数据为整数,而且$r_{ui}\in\{0,1,2,3,4,5\}$,其中0值表示用户未对此项目进行评分。

社交网络信息通常表示为有向社会关系图$G=(U,E)$,其中U是用户集合,边集合E表示用户之间的社交信任关系。$T_{uv}\in[0,1]$表示用户u和用户v之间信任权重,$T_{uv}=0$意味着用户u和用户v之间没有建立信任关系。所有用户之间的信任关系组成信任关系矩阵T。

基于社交网络推荐系统目的就是利用用户-项目评分矩阵和社交网络信息预测用户u对项目i的评分$\hat{r}_{ui}$,从而为用户推荐用户可能感兴趣的项目。

2.2 基于用户的协同过滤推荐算法

基于用户的协同过滤算法[2]假设相似的用户具有类似的偏好,具体流程为:(1)通过某种相似度度量指标找到与当前活动用户相似的用户;(2)利用相似用户对目标项目评分的权重平均来预测当前活动用户对目标项目的评分。采取合适的度量指标计算用户之间的相似度是基于用户协同过滤算法的关键。典型的相似度度量指标包括余弦相似度、皮尔逊相关系数和调整的余弦相似度。如根据余弦相似度度量指标,用户u,v之间的相似度定义为

$$\mathrm{sim}(u,v)=\cos(u,v)=\frac{\sum_i r_{ui}\times r_{vi}}{\sqrt{\sum_i r_{ui}^2}\times\sqrt{\sum_i r_{vi}^2}} \tag{1}$$

式中:i是用户u,v共同评分过的项目。

基于用户的协同过滤算法按照某种相似度度量指标计算用户之间的相似度后,找到与当前活动用户最相似的用户集合。然后,按照式(2)预测用户u对项目j的评分。

$$r_{uj}^{\mathrm{R}}=\bar{r}_u+\frac{\sum_{v\in F}\mathrm{sim}(u,v)\times(r_{vj}-\bar{r}_v)}{|\sum_{v\in F}\mathrm{sim}(u,v)|} \tag{2}$$

式中:$\bar{r}_u$表示用户u对所有已评分项目的平均评分,$\bar{r}_v$表示用户v对所有已评分项目的平均评分,F表示与用户u最相似的用户集合。

最后,基于用户的协同过滤算法将预测评分最高的K个项目推荐给当前用户。

3 基于图嵌入模型的协同过滤算法

传统协同过滤推荐算法存在严重的数据稀疏和冷启动问题。基于社交网络的协同过滤算法认为朋友之间具有共同的兴趣偏好,通常利用社交网络中用户之间的显式信任关系改进传统协同过滤算法。但是,传统的基于社交网络的协同过滤算法存在如下问题:(1)社交网络中用户之间的信任关系粒度粗,即只能用0或1来表示用户之间的信任程度;(2)仅考虑显式信任关系,而忽视隐式信任关系。显式信任关系为可观测的、具有直接连接关系的信任关系,它捕获社交网络的局部结构信息。另外,很多用户虽然没有直接建立信任关系,但是他们具有很多共同的邻居,他们之间也有可能具有较高的信任程度。即使两个用户没有建立直接的信任关系,如果具有类似的邻居结构,这两个用户之间很有可能具有较强的信任关系。在本文中,称社交网络观测到的信任关系为一阶信任度,由邻居结构推导的信任关系为二阶信任度。二阶信任关系捕获了社交网络的全局结构信息。在进行推荐模型建模时,同时考虑显式的一阶信任关系和隐式的二阶信任关系可以有效地提高推荐模型的性能。

因此,本文利用LINE模型从原始社交网络中推导同时保存社交网络局部和全局信息的用户信任关系,提出基于图嵌入模型的协同过滤推荐算法。下面首先描述LINE模型学习同时保存局部和全局信息的节点嵌入式表示的过程,然后详细解释基于图嵌入模型的协同过滤推荐算法模型。

3.1 LINE模型

作为图嵌入模型的重要代表,LINE模型适用于大规模同构信息网络,包含无向/有向、权重和无权重信息网络。另外,LINE模型在将信息网络嵌入到低维空间时,能够有效保存信息网络的全局结构和局部结构信息。其中,局部结构由网络中观测到的链接表示,捕获顶点之间的一阶信任度。 全局网络结构由顶点的共享邻域结构来确定,捕获顶点之间的二阶信任度。

LINE模型为了建模一阶信任度,对于每个无向边(u,v),定义顶点X_u和X_v之间的联合概率分布为

$$P_1(X_u,X_v)=\frac{1}{1+\exp(-\boldsymbol{Y}_u^{\mathrm{T}}\cdot\boldsymbol{Y}_v)} \tag{3}$$

式中:$\boldsymbol{Y}_u\in\mathbf{R}^{d_1}$是顶点$X_u$的低维向量表示。顶点$X_u$和$X_v$之间的经验分布为

$$\hat{P}_1(X_u,X_v)=\frac{w_{uv}}{W} \tag{4}$$

式中:$W=\sum_{(u,v)\in E}w_{uv}$,$w_{uv}$为边的权重。通过最小化信息网络中顶点之间联合概率分布和经验概率分布的KL散度来学习信息网络中顶点的低维表示,即目标函数为

$$O_1=-\sum_{(u,v)\in E}w_{uv}\log P_1(X_u,X_v) \tag{5}$$

二阶信任度意味着具有相似邻居结构两个顶点很可能具有较高程度的信任度。换句话说,每个顶点也被看作特定的“上下文”,并且假定有相同“上下文”的顶点之间存在信任关系。因此,每个顶点扮演两个角色:顶点本身和其他顶点的特定“上下文”。对于每个有向边(u,v),定义由顶点X_u生成“上下文”X_v的概率分布为

$$P_2(X_v|X_u)=\frac{\exp(\boldsymbol{Y}_v^{*\mathrm{T}}\cdot\boldsymbol{Y}_u)}{\sum_{k=1}^{|U|}\exp(\boldsymbol{Y}_v^{*\mathrm{T}}\cdot\boldsymbol{Y}_u)} \tag{6}$$

式中:$|U|$为顶点或“上下文”的数量,$\boldsymbol{Y}_v^*\in\mathbf{R}^{d_2}$为$X_v$作为“上下文”时的低维表示。由顶点$X_u$生成“上下文”$X_v$的经验分布为

$$\hat{P}_2\left(X_v|X_u\right)=\frac{w_{uv}}{d_u} \tag{7}$$

式中:d_u为顶点X_u的出度,即$d_u=\sum_{v\in N_{(u)}}w_{uv}$,$N_{(u)}$为$X_u$的邻居的集合。

类似于建模信息网络中节点的一阶信任度,为了保存信息网络的二阶信任度,通过KL散度得到目标函数为

$$O_2=-\sum_{(u,v)\in E}w_{uv}\log P_2(X_v,X_u) \tag{8}$$

LINE模型首先分别最小化目标函数O_1和O_2,学习保存一阶信任度的低维表示和保存二阶信任度的低维表示。然后合并两种低维表示作为顶点的低维特征向量,以保存信息网络的局部和全局特征。因此,对于每个顶点X_u,可以用$\boldsymbol{Y}_u\in\mathbf{R}^d$来表示,其中$d=d_1+d_2$。

3.2 基于图嵌入模型的协同过滤算法

LINE模型学习的用户节点的嵌入式表示同时捕获社交网络的局部和全局结构。基于用户节点的嵌入式表示,用户之间的信任度定义为

$$S_{uv}=\frac{\boldsymbol{Y}_u^{\mathrm{T}}\cdot\boldsymbol{Y}_v}{\left\|\boldsymbol{Y}_u\right\|_2\left\|\boldsymbol{Y}_v\right\|_2} \tag{9}$$

与社交网络中原始的信任关系相比,由式(9)得到的用户信任度不仅粒度细、密度稠,而且同时捕获了社交网络用户之间显式的一阶信任关系和隐式的二阶信任关系。换句话说,如果用户之间不存在显式信任关系,但是用户具有相似的邻居结构,根据式(9)计算的用户信任度S_{uv}捕获了用户之间的二阶信任关系。

类似于基于内存的协同过滤算法，根据式(9)计算得到用户之间的信任关系后，采用式(10)计算用户 u 对项目 j 的预测评分。

$$r_{uj}^{\mathrm{T}}=\bar{r}_u+\frac{\sum_{v\in H}S_{uv}\times(r_{vj}-\bar{r}_v)}{|\sum_{v\in H}S_{uv}|} \tag{10}$$

式中：$\bar{r}_u$ 表示用户 u 对所有已评分项目的平均评分，$\bar{r}_v$ 表示用户 u 对所有已评分项目的平均评分，r_{vj} 表示用户 v 对项目 j 的评分，H 表示用户 u 最信任的用户的集合。需要注意的是：如式(2)所示，在经典的基于用户的协同过滤算法中，$\mathrm{sim}(u,v)$ 表示根据用户评分数据计算得到的用户之间的相似度，而在式(10)中，S_{uv} 表示根据社交网络结构推导的用户信任度，这种用户信任度捕获了社交网络用户之间的显式的一阶信任关系和隐式的二阶信任关系。

用户对项目的最终评分受到相似用户和信任用户的双重影响。因此，利用权重参数平衡两方面的影响，即用户 u 对项目 j 的最终预测评分定义为

$$r_{uj}=\alpha r_{uj}^{\mathrm{R}}+(1-\alpha)r_{uj}^{\mathrm{T}} \tag{11}$$

式中：r_{uj}^{R} 表示基于评分数据相似用户的影响下，用户 u 对项目 j 的预测评分，根据式(2)计算；r_{uj}^{T} 表示基于信任用户的影响下，用户 u 对项目 j 的预测评分，根据式(10)计算。两部分评分由权重参数 α 集成，构成用户 u 对项目 j 的最终预测评分。

综上所述，基于图嵌入的协同过滤算法的框架如图1所示。

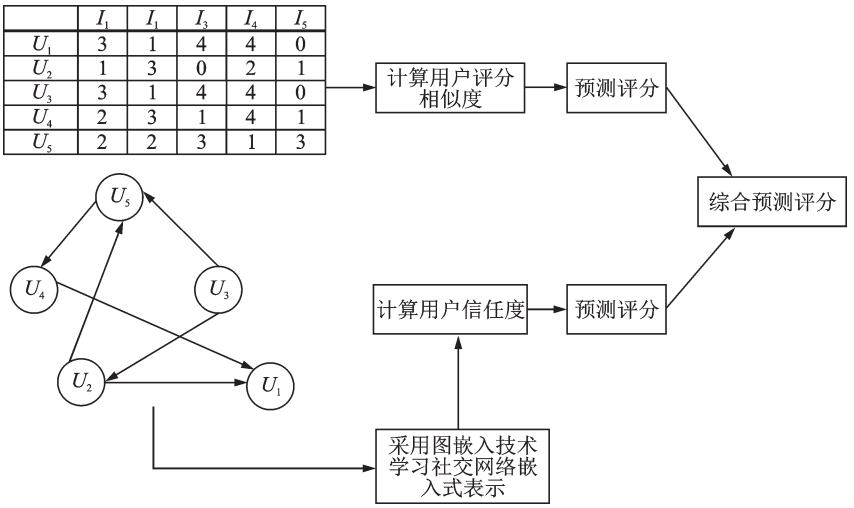


图1 基于图嵌入模型的协同过滤算法框架图

Fig.1 Framework of graph embedding model based collaborative filtering algorithm

4 实验与分析

为了验证基于图嵌入协同过滤推荐算法的性能，在真实的数据集上与其他流行的推荐算法进行了对比分析。另外，将本文提出的基于图嵌入协同过滤算法称为“GE-CF”。

4.1 数据集

选择 Epinions 数据集和 FilmTrust 数据集进行实验对比分析。Epinions 数据集中包括用户信任关

系、评分和项目类别等信息。Epinions中信任关系是有向的,信任值是二值的。文中的Epinions数据集由文献[18]的作者提供。此数据集包含922 267条评分记录,22 166个用户和296 277个产品,355 813条信任关系。FilmTrust数据集由文献[19]的作者提供。FilmTrust数据集包括1 642个用户和2 071个产品,35 497条评分记录,1 853条信任关系。

4.2 度量指标

本文采用在推荐算法中被广泛使用的评价指标:均方根误差(Root mean squared error, RMSE)和平均绝对值误差(Mean absolute error, MAE)来评价推荐算法的性能。RMSE和MAE的定义分别为

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{(i,j)\in R_{\mathrm{test}}}|r_{ij}-\hat{r}_{ij}|^2}{|R_{\mathrm{test}}|}} \tag{12}$$

$$\mathrm{MAE}=\frac{\sum_{(i,j)\in R_{\mathrm{test}}}|r_{ij}-\hat{r}_{ij}|}{|R_{\mathrm{test}}|} \tag{13}$$

式中:r_{ij}表示实际的评分值,$\hat{r}_{ij}$表示系统预测的评分值,$|R_{\mathrm{test}}|$表示测试数据集中的记录条数。RMSE和MAE值越小,推荐算法的推荐性能越好。

4.3 实验设置

为了验证本文提出推荐算法的有效性,选取如下对比方法。

UserAVG:UserAVG利用当前活动用户对所有评分项目评分的平均值作为当前活动用户对目标项目的预测评分。

UserKNN: 基于用户的协同过滤算法[2]根据用户的历史评分找到与当前活动用户兴趣相似的用户,然后利用相似用户对目标项目评分的权重平均预测当前活动用户对目标项目的评分。

PMF: PMF[13]由Mnih和Salakhutdinov提出。PMF可以看作是SVD模型的概率扩展。PMF从用户-项目评分矩阵中学习用户和项目的隐式特征向量,并使用用户和项目的隐式特征向量的内积预测用户-项目评分矩阵中缺失项。

随机将用户-项目评分矩阵分割5次,其中每次抽取80%的数据作为训练集,剩余20%的数据作为测试集。最后的运行结果取算法在5个测试集的平均值。为了公平对比,参照对比算法的相应文献或者实验结果设置不同算法的参数,在这些参数设置下,各对比算法取得最优性能。在UserKNN中,采用余弦计算用户之间的相似度;在PMF中,隐藏特征向量维度为10,正则化系数$\lambda_u=\lambda_v=0.001$;在GE-CF中,采用余弦计算用户之间的相似度。

实验运行环境为:4核Intel(R) Core(TM) i5-7400H CPU, 3.00 GHz主频,8 GB内存,Window10操作系统和J2SE1.8。

4.4 性能对比

为了过滤掉信任度较弱的信任关系,设置信任度阈值δ,用户之间信任强度小于δ的信任关系被过滤。在Epinions和FilmTrust数据集上,分别设置信任度阀值$\delta=0.95$和$\delta=0.8$,邻居数量$N=20$,LINE模型中低维向量节点的维度分别为 $d=128$ 和 $d=16$,集成参数分别为$\alpha=0.6$和0.7。所有对比算法在2个数据集上的实验结果分别如表1和表2所示。

从表1和表2可以发现:UserAVG算法性能较差,这是由于它们没有考虑物品的差异,属于非个性化的方法。除了GE-CF外,PMF的性能优于其他对比方法,再次验证了矩阵分解方法的有效性。在2

个数据集上,本文提出的基于图嵌入模型的协同过滤推荐算法的推荐准确度高于其他对比算法,验证了本文提出算法的有效性。在 Epinions 和 FilmTrust 数据集上,以 RMSE 为参考指标,与 UserAVG, UserKNN 和 PMF 中的最优结果对比,本文提出算法的改进幅度分别为 2.30% 和 1.94%。

表 1 在 Epinions 数据集上的性能对比

Table 1 Performance comparison on the Epinions dataset

Recommendation algorithm	RMSE	MAE
UserAVG	1.135 033 0	0.882 111 6
UserKNN	1.108 644 2	0.847 735 0
PMF	1.092 327 3	0.839 864 2
GE-CF	1.083 132 0	0.821 622 6

表 2 在 FilmTrust 数据集上的性能对比

Table 2 Performance comparison on the FilmTrust dataset

Recommendation algorithm	RMSE	MAE
UserAVG	0.854 695 2	0.674 076 6
UserKNN	0.843 178 2	0.663 775 4
PMF	0.835 686 8	0.651 786 6
GE-CF	0.826 834 6	0.646 383 8

4.5 参数 α 的影响

在 GE-CF 算法中,集成参数 α 对推荐性能有重要影响。α 值越大意味着由用户评分得到的预测评分比重越大,在 GE-CF 中预测评分更依赖评分数据。相反,较小的 α 值意味着在 GE-CF 中预测评分更依赖社交网络。在 Epinions 中,设置 δ 为 0.9,LINE 模型中节点低维向量维度为 d=128。在 FilmTrust 中设置 δ 为 0.4,LINE 模型中节点低维向量维度为 d=16。另外,在 Epinions 中,邻居数量 N 设置为 5;在 FilmTrust 中,N 设置为 30。在本节中,通过一组实验对集成参数 α 进行敏感度分析。实验结果如图 2 所示。

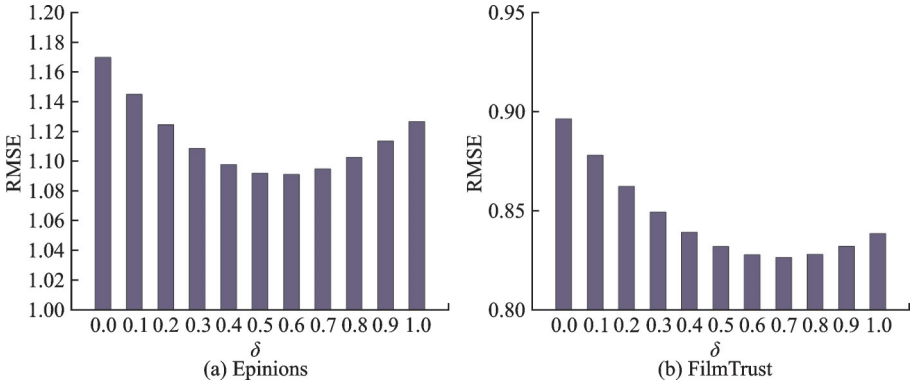


图 2 集成参数 α 对推荐性能的影响

Fig.2 Impact of parameter α on recommendation performance

从图 2 可以看出:在 2 个数据集中,随着 α 的递增,RMSE 先减小,达到最小值后,然后逐渐增加。另外,在 Epinions 中,GE-CF 在 α 为 0.6 附近时性能最优;在 FilmTrust 中,α 为 0.7 左右时,GE-CF 的性能最优。结果表明:由评分数据和社交网络数据得到的预测评分对最终的预测评分都有贡献,而且更依赖评分数据得到的预测评分。

4.6 参数 δ 的影响

信任度阈值 δ 是影响本文提出的基于图嵌入推荐算法性能的另一个重要参数。较大的 δ 值意味着过滤掉信任度较弱的用户信任关系,在 GE-CF 中融合数量少、信任度强的用户信任关系。相反,较小的 δ 值意味着在 GE-CF 中融合数量多、信任度相对小的用户信任关系。在本节中,通过一组实验对信任

度阈值δ进行敏感度分析。在Epinions中,设置δ为0.8,0.85,0.9和0.95,LINE模型中用户节点低维向量维度为$d=128$。在FilmTrust中设置δ为0.4,0.5,0.6,0.7和0.8,LINE模型中用户节点低维向量维度为$d=128$。另外,在两个数据集中,邻居数量N设置为5。实验结果如图3所示。

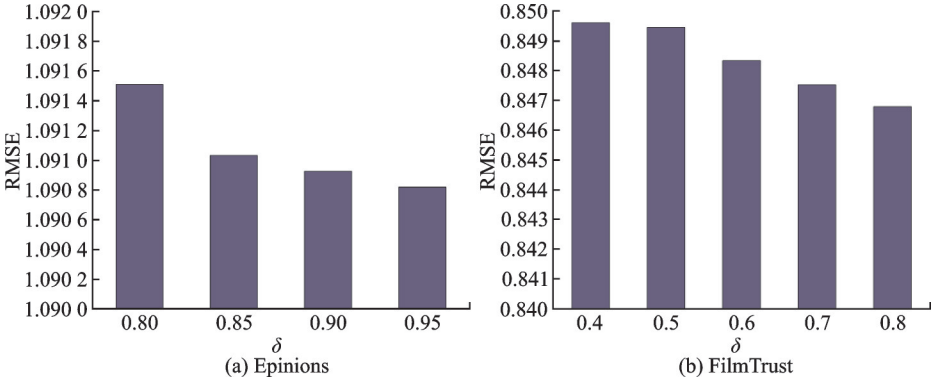


图3 参数δ对推荐性能的影响

Fig.3 Impact of parameter δ on recommendation performance

从图3可以观察到,在2个数据集中,RMSE的值大致变化趋势为:随着δ的增加,RMSE逐渐降低,推荐准确性提高。结果表明:集成细粒度、信任度强的用户信任关系更加有益于GE-CF改进推荐性能。最后,在2个数据集中,GE-CF并没有在同一个δ值下获得最低RMSE。可能的原因是:在不同的社交网络中,网络嵌入技术学习不同规模的用户信任强度。如,在Epinions中,最大用户信任度为0.975 3,而在FilmTrust中,最大信任度为0.883 9。

4.7 参数d的影响

在本节中,改变d的值,观察2个数据集上RMSE的变化情况。在Epinions中,设置$\delta=0.9$,$N=5$,d从64变化到256;由于FilmTrust数据集规模较小,增大d时,发现用户之间的最大信任度减小很快。因此,在不同d的情况下设置了不同的最大δ值,且设置$N=5$,d从16变化到256。实验结果分别如表3和表4所示。

表3 参数d的影响(Epinions)

Table 3 Impact of parameter d (Epinions)

The value of d	RMSE
64	1.090 845 4 ($\delta=0.95$)
128	1.090 818 6 ($\delta=0.95$)
256	1.090 816 6 ($\delta=0.90$)

表4 参数d的影响(FilmTrust)

Table 4 Impact of parameter d (FilmTrust)

The value of d	RMSE
16	0.846 787 0 ($\delta=0.8$)
32	0.848 498 8 ($\delta=0.7$)
64	0.849 906 0 ($\delta=0.5$)
128	0.849 454 8 ($\delta=0.5$)
256	0.851 675 2 ($\delta=0.5$)

从表3和4可以观察到,在Epinions数据集上,$d=256$时,GE-CF性能最优;在FilmTrust上,$d=16$时,GE-CF的性能最优。另外,随着d的增加,较难挖掘到较强信任度的用户信任关系。可能的原因是:随着d的增加,LINE模型学习到用户嵌入式表示可以表示较多的隐式特征,但同时会引入一些噪声,降低用户之间信任度计算的准确性。

5 结束语

在本文中,提出了一种基于图嵌入模型的协同过滤推荐算法,利用图嵌入技术学习社交网络中用户的低维嵌入式表示,并根据用户的嵌入式表示推导用户之间细粒度的信任关系。这种细粒度的信任关系捕获了用户之间的一阶信任度和二阶信任度。最后,以信任用户和相似用户对目标物品的评分权重预测用户对目标物品的评分。在真实数据集上的实验结果证明,基于图嵌入模型的协同过滤算法的性能优于传统的协同过滤算法。本文仅考虑在基于内存的协同过滤算法中融合图嵌入技术,在矩阵分解模型中集成图嵌入技术,以改进传统协同过滤算法的性能是未来的研究方向。

参考文献:

[1] 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 (6): 734-749.

[2] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. [S.l.]: [s.n.], 1998: 43-52.

[3] LINDEN G, SMITH B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003 (1): 76-80.

[4] 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 (CIKM′08). [S.l.]: ACM, 2008: 931-940.

[5] MA H, KING I, LYU M R. Learning to recommend with social trust ensemble[C]//Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR′09). [S.l.]: ACM, 2009: 203-210.

[6] JAMALI M, ESTER M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys′10). [S.l.]: ACM, 2010: 135-142.

[7] YANG B, LEI Y, LIU D, et al. Social collaborative filtering by trust[C]//Proceedings of the Twenty-Third international Joint Conference on Artificial Intelligence (AAAI′13). [S.l.]: [s.n.], 2013: 2747-2753.

[8] CAI Hongyun. A comprehensive survey of graph embedding: Problems, techniques, and applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9):1616-1637.

[9] 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. [S.l.]: ACM, 2014: 701-710.

[10] AHMED A, SHERVASHIDZE N, NARAYANAMURTHY S, et al. Distributed large-scale natural graph factorization[C]// Proceedings of the 22nd International Conference on World Wide Web. [S.l.]: [s.n.], 2013: 37-48.

[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. [S.l.]: [s.n.], 2015: 1067-1077.

[12] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42 (8): 30-37.

[13] MNIH A, SALAKHUTDINOV R. Probabilistic matrix factorization[C]//Proceedings of the 21st Annual Conference on Neural Information Processing Systems. [S.l.]: [s.n.], 2007: 1257-1264.

[14] 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 (KDD′08). [S.l.]: ACM, 2008: 426-434.

[15] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.

[16] WEIMER M, KARATZOGLOU A, LE Q V, et al. Maximum margin matrix factorization for collaborative ranking[C]// Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS′07). [S.l.]: [s.n.], 2007:1-8.

[17] YU Kai, ZHU Shenghuo, LAFFERTY J, et al. Fast nonparametric matrix factorization for large-scale collaborative filtering

[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR′09). [S.l.]: ACM, 2009: 211-218.

[18] TANG J, HU X, GAO H, et al. Exploiting local and global social context for recommendation[C]//Proceedings of International Joint Conference on Artificial Intelligence. [S.l.]: [s.n.], 2013: 2712-2718.

[19] GOLBECK J, HENDLER J A. FilmTrust: Movie recommendations using trust in web-based social networks[C]//Proceedings of Consumer Communications and Networking Conference. [S.l.]: [s.n.], 2006: 282-286.

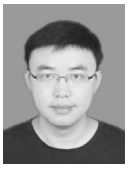
作者简介:

高海燕(1982-),讲师,研究方向:数据挖掘、用户行为分析,E-mail: gaohy@njupt.edu.cn。

毛林(1997-),男,本科生,研究方向:数据挖掘、推荐算法,E-mail: 13852499043@163.com。

窦凯奇(1998-),男,本科生,研究方向:数据挖掘、推荐算法,E-mail: jdhxdzg@163.com。

倪文烨(1998-),女,本科生,研究方向:数据挖掘、推荐算法,E-mail: 1256984540@qq.com。



赵卫滨(1979-),博士研究生,讲师,研究方向:智能化计算,E-mail: zhaowb@njupt.edu.cn。

余永红(1978-),**通信作者**,男,博士,副教授,研究方向:推荐算法和社交网络分析,E-mail: yuyh@njupt.edu.cn。

(编辑:夏道家)