

基于多关系社交网络的协同过滤推荐算法

宾 晟 孙更新

(青岛大学数据科学与软件工程学院 山东 青岛 266071)

摘 要 推荐系统是大数据中最常见的应用之一,传统的协同过滤推荐算法直接基于用户-项目评分矩阵,对于海量的用户和商品数据,算法的执行效率将会显著降低。针对这一问题,提出了一种基于多关系社交网络的协同过滤推荐算法。该算法利用信息传播方法对基于多子网复合复杂网络模型构建的多关系社交网络进行社团结构划分,从而将相似度接近的用户划分到一个社团中,进而在社团内部选择用户的 k-近邻集合来构建用户-项目评分矩阵,然后利用协同过滤算法进行推荐,从而实现了在不降低推荐准确率的前提下提升推荐算法的执行效率。在真实数据集 Epinions 上,将所提算法与传统的协同过滤推荐算法进行对比。实验结果表明,所提算法具有较高的推荐效率和准确率,特别是对于海量数据,推荐算法的执行时间缩短到原有的 1/10。

关键词 推荐算法,社交网络,大数据,多子网复合复杂网络,信息传播,社团结构

中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/jsjxx.181102189

Collaborative Filtering Recommendation Algorithm Based on Multi-relationship Social Network

BIN Sheng SUN Geng-xin

(School of Data Science and Software Engineering, Qingdao University, Qingdao, Shandong 266071, China)

Abstract Recommendation system is one of the most common applications in big data. Traditional collaborative filtering recommendation algorithm is directly based on user-item scoring matrix. For massive user and commodity data, the efficiency of the algorithm will be significantly reduced. Aiming at this problem, this paper proposed a collaborative filtering recommendation algorithm based on multi-relational social network. The information propagation method is used to detect communities in the multi-relationship social network based on multi-subnet composite complex network model, the users with similarity are divided into the same community. And then the k-nearest neighbor set of users is selected to construct the user-item scoring matrix within the community. Then the collaborative filtering algorithm is used to recommend through the new user-item scoring matrix, thus improving the efficiency of recommendation algorithm without reducing the accuracy of recommendation. Compared with traditional collaborative filtering recommendation algorithm on real data set Epinions, the results show that the proposed algorithm has high recommendation efficiency and accuracy. Especially for big data, the execution time of the proposed recommendation algorithm is improved by more than 10 times.

Keywords Recommendation algorithm, Social network, Big data, Multi-subnet composite complex network model, Information propagation, Community structure

1 引言

推荐系统是解决“信息过载”问题^[1]的有效手段,也是大数据技术的最常见应用之一。推荐系统利用知识发现技术,根据用户的历史信息、兴趣爱好等特征,筛选用户感兴趣的信息、产品,从而实现个性化推荐。目前,最常用的推荐系统技术是基于协同过滤的推荐算法^[2]。这类算法基于用户-项目评分矩阵来计算用户相似度或商品相似度,进而为用户推荐其可能需要的商品。但随着用户和商品数量的增加,协同过

滤算法的性能不断降低,这是因为大量无关商品与用户间的相似度计算耗费了大量时间,但对最终的推荐结果并没有显著影响,这也正是大数据“低价值密度”特性的体现。

此外,大数据的真正要义不在于大,而在于多源^[3]。如何在多源数据的融合过程中,通过科学地分析找出其内在的联系,从中提取出统一的、比单一数据源更丰富且更有价值的信息,才是使用大数据的真正意义。因此,将用户间的社交关系融入到推荐系统中,可以更准确地计算用户相似度,并且能在一定程度上解决传统协同过滤推荐算法中存在的数据稀疏、

到稿日期:2018-11-27 返修日期:2019-05-02 本文受教育部人文社会科学研究青年项目(15YJC860001),全国统计科学研究重点项目(2017LZ38),山东省自然科学基金面上项目(ZR2017MG011),山东省社会科学规划项目(17CHLJ16)资助。

宾 晟(1979—),女,博士,副教授,主要研究方向为大数据分析,E-mail:binsheng@qdu.edu.cn(通信作者);孙更新(1978—),男,博士,副教授,主要研究方向为复杂网络。

冷启动等问题^[4]。但社交网络中的用户间存在多种关系,每种关系对于推荐的影响是不同的,单纯引入某一种社交关系必然影响推荐结果的准确率。

因此,本文提出了一种基于多关系社交网络的协同过滤算法。首先通过多子网复合复杂网络模型^[5]构建多关系社交网络;然后利用信息传播方法对构建的多关系社交网络复合网进行社团结构划分,从而将相似度接近的用户划分到一个社团中;最后在社团内部选择与被推荐用户相似度最高的 k 个用户组成其 k -近邻集合,使用该近邻集合构建用户-项目评分矩阵,利用协同过滤算法进行推荐。由于社团结构之间存在关系稀疏的特性,使得不同社团结构的成员间的相似性较弱,因此在推荐时不必分析由整体数据集构建的用户-项目评分矩阵,从而实现了在不降低推荐准确率的前提下提升推荐算法的执行效率。

2 相关工作

协同过滤技术主要分为基于记忆^[6]的协同过滤和基于模型^[7]的协同过滤,这两类方法都是以用户-商品评分矩阵为基础的,但随着用户和商品数量的增加,用户-商品评分矩阵将变得越来越庞大,从而导致协同过滤算法的性能不断降低。为了提高推荐算法的执行效率,Roh 等^[8]利用层次聚类算法对用户和商品进行聚类,根据根节点到叶节点的加权求和进行推荐。Zhang 等^[9]使用模糊聚类方法划分喜欢相似物品的用户簇和被相同用户喜欢的物品簇,然后进行推荐。Santhini 等^[10]提出了通过聚类方法对大数据进行协同过滤推荐的基本思路,在一定程度上提高了推荐算法的效率。但是聚类技术依靠用户-商品评分矩阵作为原始数据源来反映个人偏好,没有考虑用户之间的朋友关系、信任关系等社会关系,也没有考虑被评分的商品或项目之间存在的隐性关系,忽略这些重要关系难免会造成推荐结果的失真。因此,将社交网络引入推荐系统中来完善评分矩阵的社会化推荐系统^[11]应运而生。Chang 等^[12]在推荐算法中融合了社交网络中的信任关系,将用户间的信任度作为用户-项目矩阵中空缺值的替代来预测评分。Pan 等^[13]利用社交网络中的信任关系来缓解推荐系统中的数据稀疏和冷启动问题,证明了社交网络能够在很大程度上提升推荐系统的性能。社交网络是典型的复杂网络^[14],其具有的社团结构性性质可以将网络成员划分为组内关系紧密、组间关系稀疏的社团^[15]。Wang 等^[16]利用加权标签传播算法对论文引用合作关系网络进行社团结构划分,进而在同一社团内使用推荐算法进行学术论文推荐。Zhang 等^[17]将用户-项目评分矩阵映射到用户相似性网络,结合网络中的社团将排序后的 Top- n 商品推荐给用户,提升了推荐的准确性和效率。Guo 等^[18]提出了一种融合社交网络特征的协同过滤推荐算法 TDSRec,在一定程度上解决了传统协同过滤算法中的数据稀疏性问题。Forsati 等^[19]提出的 SocialMF 算法将信任传递机制引入社交推荐中,较好地反映了社交网络信任关系对推荐的影响。Park 等^[20]提出的 TDSRec 算法则进一步将社交网络中的信任关系分为显式信任和非信任数据进行社交推荐。虽然现有的基于复杂网络社团划分的推荐算法已经在推荐的准确率和算法的执行效率方面取得了

一定的进步,但这些算法都是基于单关系社交网络的,即认为社交网络中仅仅存在一种关系,而现实中的社交网络实际上是一个多关系网络,例如社交网络的用户间存在着用户信任关系、兴趣相似关系等,这些显式或隐式关系共同决定了用户间的交互特征,从而对推荐结果产生不同的影响。因此,引入推荐系统中的社交网络应该是一个多关系的、存在异质节点的复杂网络,而传统的复杂网络模型并不能很好地描述这些错综复杂的关系,也无法表述异质节点间的相互作用。基于此,本文基于能够描述不同类个体间多关系的多子网复合复杂网络模型来构建多关系社交网络,并利用信息传播方法对构建的网络进行社团结构划分。在预测用户评分时,结合用户评分偏好程度计算用户评分相似性,综合修正后的用户评分相似性和用户社交相似性,在社团结构中为被推荐用户选择其 k -近邻集合,从而大大减小了推荐算法的计算量。

3 基于多子网复合复杂网络模型的多关系社交网络构建

在推荐系统中,其元素主要可以分为用户和商品两个集合,假设共有 m 个用户和 n 件商品,则用户集合 U 可表示为 $U=\{u_1,u_2,\cdots,u_m\}$,商品集合 I 可表示为 $I=\{i_1,i_2,\cdots,i_n\}$ 。在此基础上,用户对于商品的评分矩阵 $\mathbf{R}$ 由 $m \times n$ 个元素 $r_{11},r_{12},\cdots,r_{mn}$ 组成,表示为 $[r_{ui}]_{m \times n}$,其中 r_{ui} 表示用户 u 对于第 i 件商品的评分,用户-商品评分矩阵如表 1 所列。

	i_1	i_2	i_3	i_4	i_5
u_1	3		5		4
u_2		2		3	
u_3			2		1
u_4		4			1
u_5	2	5			4

图 1 用户-商品评分矩阵

Fig. 1 User-item scoring matrix

现实中的推荐过程常常伴随着社交信息的融合,例如用户询问好友购物意见的行为或好友购物信息的共享等,实际上若用户之间存在着信任关系,那么他们的兴趣偏好相似性会明显高于陌生人^[21],如果其中一个用户对某个商品做出了高评分,则可以认为另一个用户有很大的可能性会对该商品给出高评分。

社交网络中的这种信任关系可以用信任评分矩阵 $\mathbf{T}$ 表示, $\mathbf{T}=[t_{uv}]_{m \times m}$,其中 t_{uv} 表示用户 u 对用户 v 的评分,反映了用户之间的信任程度,信任评分矩阵如表 2 所列。

	u_1	u_2	u_3	u_4	u_5
u_1			1		
u_2					3
u_3		1			
u_4					2
u_5	2				

图 2 信任评分矩阵

Fig. 2 Trust-scoring matrix

为了有效融合这种社交特征,基于多子网复合复杂网络模型将用户商品评分网络和用户信任关系网络复合成多关系社交网络。

3.1 构建用户商品评分网络和用户信任关系网络

多子网复合复杂网络模型(简称复合网)是一个四元组 $G=(V,E,R,F)$,其中: $V=\{v_1,v_2,\cdots,v_m\}$ 表示网络中节点的集合, $m=|V|$ 是集合中节点的个数; $E=\{\langle v_h,v_l\rangle|v_h,v_l\in V,1\leq h,l\leq m\}\subseteq V\times V$ 表示节点间连边的集合; $R=R_1\times\cdots\times R_l\times\cdots\times R_n=\{(r_1,\cdots,r_l,\cdots,r_n)|r_i\in R_i,1\leq i\leq n\}$, R_i 表示节点间一种相互作用关系的集合, r_i 表示节点间的一种相互作用关系, n 是节点间相互作用关系的总数,若 $|R|\geq 2$ 则表示网络为多关系网络;映射 $F:E\rightarrow R$ 是边集 E 经过 φ 函数投影在 F 中找到唯一对应的映射,表示边上具有的关系类型。

单关系复杂网络是复合网的一种特例,利用复合网模型分别基于用户-商品评分矩阵和信任评分矩阵构建用户商品评分网络 R_G 和用户信任关系网络 T_G ,并设定阈值 R 和 T 。若用户 u 对商品 i 的评分大于或等于阈值 R ,则认为该商品满足用户偏好,类似地,若用户 u 对另一用户 v 的信任评分大于或等于阈值 T ,则认为用户之间存在信任关系。在用户商品评分网络 R_G 中,每个节点代表一个用户或一件商品, $R=R_1=\{r_1\}$ 代表用户对商品的偏好关系,在用户信任关系网络 T_G 中,每个节点代表一个用户, $R=R_2=\{r_2\}$ 代表用户之间的信任关系。

根据图 1 和图 2 给出的用户-商品评分矩阵和用户信任评分矩阵构建网络,其中用户商品评分阈值 R 和用户信任评分阈值 T 都设置为 2,则构建的网络拓扑结构分别如图 3 和图 4 所示。

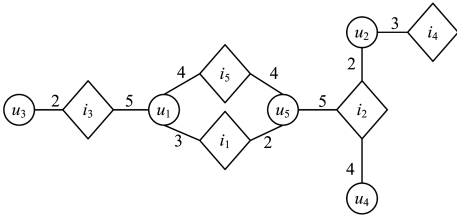


图 3 用户商品评分网络
Fig. 3 User-item scoring network

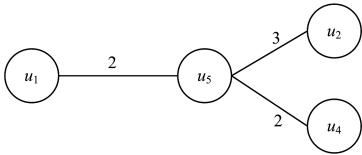


图 4 用户信任关系网络
Fig. 4 User trust relationship network

3.2 构建多关系复合网络

复合网能通过子网加载运算复合两个或两个以上的复杂网络,在加载过程中,将网络映射成向量空间,节点的相互关系映射成向量空间中的多维向量,因此子网加载过程可以转化成向量空间的基变换,相关节点间相互关系的变化通过空间向量的相关运算完成。构建的多关系复合网由于继承了每个子网的性质,其关系向量空间维数自然增大。

通过子网加载运算将用户评分网络和用户信任关系网络加载成一个多关系复合网。子网加载过程如下:设向量复合网 $\Sigma_1=(G_1,S_1,M_1),\Sigma_2=(G_2,S_2,M_2)$,其中 $G_1=(V_1,E_1,R_1,F_1),G_2=(V_2,E_2,R_2,F_2)$,设关系集合为 R' , G_1 为基底网,在加载映射 $\Psi:V_1\times V_2\rightarrow 2^{R'}$ 与关系强度映射 $\Psi:(V_1\times V_2)\times R'\rightarrow \bigcup_{r'\in R'}\text{dom}(r')$ 下,将 Σ_2 加载到 Σ_1 是指产生一个新的复合网 $\Sigma=(G,S,M)$,其中 $G=(V,E,R,F)$,称 R' 为加载关系集合, $\langle v_h,v_l\rangle\in V_1\times V_2$ 为外部边, v_h,v_l 为边界节点。

加载完成后的复合网的拓扑性质相比传统的复杂网络有所区别,其与节点关联的边包含了所有子网的关系,因此定义复合网中节点的混合权和混合度。

节点混合权:设节点 $v_h\in V$,关系强度比例系数为 $SF=(sf_1,sf_2,\cdots,sf_k)$,则节点 v_h 的混合权 $k_{v_h}^R$ 为:
$$k_{v_h}^R = \sum_{r_i\in R} sf_i \times k_{v_h}^{r_i}$$

其中, $k_{v_h}^{r_i}$ 为节点 v_h 关于关系 r_i 的权。

节点混合度:设节点 $v_h\in V$,关系强度比例系数为 $SF=(sf_1,sf_2,\cdots,sf_k)$,则节点 v_h 的混合度 $\hat{k}_{v_h}^R$ 为:
$$\hat{k}_{v_h}^R = \sum_{r_i\in R} sf_i \times \hat{k}_{v_h}^{r_i}$$

其中, $\hat{k}_{v_h}^{r_i}$ 为节点 v_h 关于关系 r_i 的度。

将图 3 所示的用户商品评分网络 R_G 和图 4 所示的用户信任关系网络 T_G 加载成复合推荐网络,设关系 r_1 和 r_2 的关系强度参数 sf_1 和 sf_2 均为 1,子网 R_G 的边权集合为 $\{1\}$,子网 T_G 的边权集合为 $\{0.5\}$,则构建的复合网如图 5 所示。

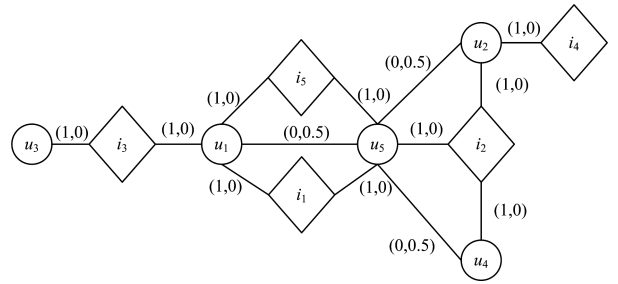


图 5 复合推荐网络
Fig. 5 Composite recommendation network

图 5 中,节点 u_5 关于关系 r_1 的度 $\hat{k}_{u_5}^{r_1}=3$,关于关系 r_2 的度 $\hat{k}_{u_5}^{r_2}=3$,则节点 u_5 的混合度为 $\hat{k}_{u_5}^R=1\times\hat{k}_{u_5}^{r_1}+1\times\hat{k}_{u_5}^{r_2}=6$;节点 u_5 关于关系 r_1 的权 $k_{u_5}^{r_1}=1+1+1=3$,关于关系 r_2 的权 $k_{u_5}^{r_2}=0.5+0.5+0.5=1.5$,则节点 u_5 的混合权为 $k_{u_5}^R=1\times k_{u_5}^{r_1}+1\times k_{u_5}^{r_2}=4.5$ 。

4 多关系社交网络中的社团结构划分

随着用户和商品数量的增加,用户做出过评分的物品将非常有限,这使得用户-商品评分矩阵中出现大量空缺值,基于该矩阵进行相似度计算就会出现很多相似度为零的情况;此外,推荐算法的性能也将随着用户和商品数量的增加而降低。为了缓解这两个问题,可以将推荐系统中的对象进行社团结构划分,社团结构之间关系稀疏的特性使得不同社团成

员间的关联性较弱,推荐时不必分析整体数据集构建的用户-商品评分矩阵,仅考虑被推荐用户所在社团的内部成员即可,这将极大地降低计算量和数据稀疏度。随后通过综合评分偏好程度和社交相似性等特征来补全评分,使推荐结果的准确率并不会因为用户-商品评分矩阵规模的减小或者数据的筛选而降低。但传统的社团划分方法并不适用于多关系网络,为此本文采用基于信息传播的多关系网络社团结构发现算法对推荐复合网进行社团结构划分。

4.1 多关系复杂网络上的信息传播过程

初始时网络中的所有节点都没有携带信息,首次传播开始,设置其中一个节点*i*为信源节点并赋予其非空信息量*I*,设该节点的邻居节点集合为*Vⁱ*,此时网络中的信息分布为*X*=(0,⋯,*I*,⋯,0),信源节点通过连边向*Vⁱ*中的节点传播信息,其邻居节点接收完信息后将继续向各自的相邻节点集合*Vⁱⁱ*传播信息,但作为接收方的节点并不是完全接收来自信源节点的信息,而是以一定的概率*p*接收;第二次传播时,集合*Vⁱ*=(0,⋯,*I*₁,⋯,0,⋯,*I*_{*j*},⋯,0,⋯,*I*_{*n*})表示所有携带信息的节点集合,集合*Vⁱ*中的节点既可以向邻居节点传递信息,也可以接收来自其相邻节点的信息。由于接收方节点并不能接收全部的外来信息,当网络规模较大时,仅靠单次传播,网络中部分距离信源节点较远的节点接收到的信息较少,不能充分表示信源节点对这些节点的影响,因此只有信息传播在网络中迭代*T*次后,网络中所有节点携带的信息量分布情况才可以代表信源节点对整个网络中信息传播的影响能力。根据每个节点作为信源节点时网络中信息量的分布向量,可以计算网络节点间的相似性。

我们将信息传播扩展到多关系社交网络中,*R*=(*r*₁,*r*₂,⋯,*r*_{*i*},⋯,*r*_{*n*})表示网络中存在的关系集合,基于多子网复合复杂网络模型构建包含*n*种关系的复合网。与单关系复杂网络类似,复合网中的每个节点也是按照一定的比例*p*接收信息,但不同的是需要考虑节点与邻居节点连边上所具有的各种关系的相互作用对接收信息概率*p*的影响。

复合网信息传播概率矩阵[^]*P*的定义如下:
[^]*P*=[^]*D*⁻¹[^]*W* (1)

其中,[^]*D*为节点混合度组成的对角矩阵,[^]*W*为网络的邻接混合权矩阵。该矩阵中的元素可表示为:

[^]*W*_{*v_h,v_l*} = ∑_{*v_h,v_l*} ∈ *V* *s**f*_{*r_i*} × *k*_{*r_i*}' (2)

邻接混合权矩阵中的元素是对节点*v_h*和节点*v_l*间所存在的各种关系的关系强度的加权求和。其中,*s**f*_{*r_i*}表示节点*v_h*和节点*v_l*之间关于关系*r_i*的关系强度比例系数,*k*_{*r_i*}'为关系*r_i*的关系强度。

在邻接混合权矩阵的基础上,节点*v_h*的混合权也可表示为:

*k*_{*v_h*}^{*R*} = ∑_{*v_l*} ∈ *V* [^]*W*_{*v_h,v_l*} (3)

依次选择网络中的节点作为信源节点并赋予1个单位的信息量,网络中其余节点的信息量为0。以网络中一个节点*s*为例,设节点*s*为信源节点,其携带的信息量为1,初始化节点*s*表示的向量为*X*⁰=(0,0,⋯,1,0)。第一次传递过程中由该

节点向其邻居节点传递信息,这一过程用数学公式表示为*X*_{*s*}¹=[^]*PX*_{*s*}⁰,向量*X*_{*s*}¹中的元素为第一次传递后网络中各个节点所携带的信息量;第二次传递过程中,接收到信息的节点向其相邻节点传递信息,若其相邻节点也具有信息,其也可以接收来自相邻节点的信息,这一过程的数学表达式为*X*_{*s*}²=[^]*PX*_{*s*}¹,向量*X*_{*s*}²中的元素为第二次传递后网络中各个节点携带的信息量;类似地,每次传播后的信息表示为*X*_{*s*}^{*t*+1}=[^]*PX*_{*s*}^{*t*}。在下次传播前,设*X*_{*s*}^{*t*+1}=1,即信源节点在*t*+1次传播开始时信息量始终为1。这样网络中各节点经过*T*次传播后都将得到各自的信息量分布向量,利用这些向量将得到能表示节点影响力的信息分布矩阵。

以图5中的复合网为例,其概率传播矩阵[^]*P*为:

[^]*P*=

4	0	0	0	0	0	0	0	0	0
0	3	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0
0	0	0	0	6	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	3	0	0	0
0	0	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	2

×

0	0	0	0	0.5	1	0	1	0	1
0	0	0	0	0.5	0	1	0	1	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.5	0	1	0	0	0
0.5	0.5	0	0.5	0	1	1	0	0	1
1	0	0	0	1	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0

分别以节点*u*₁、节点*u*₂和节点*i*₅为源节点进行信息传播。首次传播得到的向量如下:

*X*₁=[0,0,0,0,0.13,0.25,0,0.25,0,0.25]
*X*₂=[0,0,0,0,0.17,0.33,0,0.33,0]
*X*₃=[0.5,0,0,0,0.5,0,0,0,0,0]
第二次传播得到的向量如下:
*X*₁=[0.39,0.01,0.13,0.01,0.38,0.27,0.02,0.25,0,0.27]
*X*₂=[0.01,0.46,0,0.13,0.28,0.03,0.36,0,0.33,0.03]
*X*₃=[0.54,0.04,0,0.04,0.56,0.21,0.08,0.13,0,0.21]

4.2 基于信息传播的多关系社交网络社团结构发现算法

将网络中的节点用向量表示后,节点就成为了适合聚类算法处理的数据结构,并且依据网络中的信息传播计算网络节点间的相似性,从而将成熟的数据挖掘聚类算法应用于多关系社交网络的社团结构发现。

前期研究基于经典 K-means 聚类算法提出了 CSDM 算法来发现多关系社交网络中的社团结构^[15]。K-means 算法的最大缺点是对初始聚类中心点的选择敏感,聚类中心点的初始值将直接影响最终聚类的结果。

在复杂网络中通常把节点的重要性称为中心性,作为网络中刻画节点重要性的基础指标,度中心性认为节点的影响力与邻居节点的数量有关,即邻居节点数目越多,该节点在网络中的影响力就越大。设节点 v_h 与节点 v_l 关于关系 r_i 的路径长度为 $l_{v_h v_l}^{r_i}$,定义节点 v_h 关于关系 r_i 的中心度为:

$$c_{v_h}^{r_i} = \frac{\min_{v_g \in V} \sum_{v_l \in V} l_{v_h v_l}^{r_i}}{\sum_{v_l \in V} l_{v_h v_l}^{r_i}}$$

其中, $\min_{v_g \in V} \sum_{v_l \in V} l_{v_h v_l}^{r_i}$ 表示节点 v_h 与其他所有节点关于关系 r_i 的路径长度之和的最小值, $\sum_{v_l \in V} l_{v_h v_l}^{r_i}$ 表示节点 v_h 与其他节点关于关系 r_i 的路径长度之和。中心度 $c_{v_h}^{r_i}$ 的值越接近于 1,说明该节点与其他所有节点的路径长度之和越小,其中心地位越强;中心度的值越接近于 0,说明其中心地位越弱。

CSDM 算法将选取复合网中中心度较大的节点作为初始聚类中心点,然后按照经典 K-means 聚类算法的步骤,对多关系社交网络进行社团结构划分。

5 基于社团结构划分的协同过滤推荐算法

通过网络中的信息传播解决了推荐系统用户组成的社交网络由于多关系而难以计算相似性的问题,传播完成后各个成员携带的信息量反映了不同节点在网络中的相似程度。将相似度很高的用户划分到一个社团中,可使得社团成员有共同的评分偏好和兴趣偏好等。从社团结构中包含的用户信任关系可以得到社交偏好向量,该向量代表了用户的社交倾向,这种倾向可以用社交相似度来衡量;从社团结构中包含的评分信息可以得到用户的兴趣偏好,通过提取的兴趣偏好向量可以确定兴趣偏好相似度。以一定比例综合社交相似度和兴趣偏好相似度,可得到一个新的综合相似度指标——推荐度,利用该指标可以进行商品的排序和推荐,具体的推荐流程如图 6 所示。

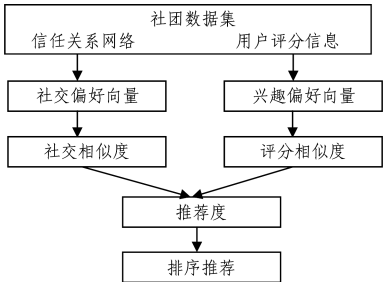


图 6 基于社团结构的推荐流程

Fig. 6 Flow of recommendation based on community

5.1 用户社交相似度和评分相似度

假设用户 u 和用户 v 存在社交关系,则认为 u 和 v 之间的兴趣更加接近,其偏好相似程度高于陌生人^[22],而这种具有促进推荐效果的社交关系可以表现为用户间的信任关系。

本文将同一个社团中的用户映射成新的用户信任关系网络,再利用信息传播法将网络中的用户映射成向量,用 Pearson 相似度衡量用户 u 和用户 v 的社交相似度,具体公式如下:

$$sim_{social}(u, v) = \frac{\sum_l (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_l (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_l (r_{vi} - \bar{r}_v)^2}}$$

用户对于物品的共同评分项、评分趋势等数据在一定程度上也反映了用户的兴趣爱好。本文除了以 Pearson 相关性来度量用户的评分倾向,还定义了用户偏好程度,评分倾向越一致,说明用户的兴趣偏好越接近,具体公式如下:

$$P(u, v) = \left(\frac{|I_u \cap I_v|^{(>R)}}{|I_u|^{(>R)}} + \frac{|I_u \cap I_v|^{(<R)}}{|I_u|^{(<R)}} \right) \times \left(\frac{|I_u \cap I_v|^{(>R)}}{|I_v|^{(>R)}} + \frac{|I_u \cap I_v|^{(<R)}}{|I_v|^{(<R)}} \right)$$

其中, R 为评分阈值,其可以与构建用户-商品评分网络时的阈值一致,3.1 节中用户对商品的评分低于阈值 R 表示该用户对这件商品不存在偏好关系,即不喜欢这件商品,本节用其来表示用户的评分为积极评分或者消极评分,评分大于阈值 R 表示此评分属于积极评分; I_u 和 I_v 分别表示用户 u 和用户 v 做出评分的商品集合。将用户偏好程度融入到计算评分相似性的过程中,用户对社团中的所有商品的评分构成了新的评分向量, $sim(u, v)$ 为用户 u 评分向量和用户 v 评分向量的 Pearson 相关性,这样将得到基于评分偏好的用户评分相似度:

$$sim_{rating}(u, v) = P(u, v) \times sim(u, v)$$

为了提高用户相似度的可靠性,本文采用下式融合用户社交相似度和用户评分相似度:

$$sim(u, v) = (1 - \alpha) * sim_{social} + \alpha * sim_{rating}$$

其中,参数 α 为控制两种相似度在推荐过程中所占比重的权, sim_{social} 为用户 u 和用户 v 的社交相似度, sim_{rating} 为用户 u 和用户 v 的评分相似度。当 $\alpha > 0.5$ 时,评分相似度在最终的相似度中所占比重较大,当 $\alpha < 0.5$ 时,以社交相似度为主。

5.2 预测评分

推荐系统主要依靠对被推荐用户未打分的商品进行评分预测做出推荐,最终将 Top- n 件打分最高的商品推荐给用户。预测时不需要用到社团中的所有用户,通常选择与当前用户相似度最高的 k 个近邻集合即可,在选择近邻用户的过程中,若 k 取值过大,加入预测的相似度不高的邻居会给预测带来额外的噪声;若 k 取值过小,则会因用来预测的最近邻不够导致偏差。用户 u 对商品 i 的预测评分通过下式计算:

$$r_{ui} = \bar{r}_u + \frac{\sum_{v \in S(u)} (r_{vi} - \bar{r}_v) \times sim(u, v)}{\sum_{v \in S(u)} sim(u, v)}$$

其中, S_u 是社团中与用户 u 相似度最高的 k 个用户的集合, r_{ui} 为集合 S_u 中的用户 v 对商品 i 的评分, $\bar{r}_v$ 为用户 v 所有评分商品的评分均值。

6 实验及结果分析

通过与不同经典推荐算法进行比较,验证了本文所提算法的时间效率和推荐质量,并且分析了参数设置对推荐结果的影响。

6.1 实验环境及数据集

为了对大数据集进行分析,搭建了由 5 个节点机构成的

Hadoop 集群平台,各节点机均为 Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz,8GB 内存,64 位操作系统。

实验中采用推荐系统中常用的 Epinions 数据集^[23]和 FilmTrust 数据集^[24],其中 Epinions 数据集包含用户的评分信息和社交信息。整个数据集中共包含 40 163 名用户和 139 738 件商品,评分记录达到了 664 824 条,用户对商品的评分利用数值 1 到 5 来表示,若用户之间存在信任关系,则使用评分 1 表示,若不存在信任关系,则标记为 0。数据集中的其他统计信息如表 1 所列。

表 1 Epinions 数据集的统计信息
Table 1 Statistics of Epinions dataset

名称	最小评分数量	最大评分数量	平均评分数量
用户	1	1022	16.55
商品	1	2018	4.76

为了使用本文所提算法进行评分预测,需要首先确定评分关系和信任关系的关系系数和关系强度,为此选取编号为 2 和 10 004 的用户作为种子节点,剔除某些明显异常值后,共选取了 150 个样本数据作为计算两种关系对应的关系比例系数的线性回归分析数据样本集。针对实验数据中的多关系社交网络中存在的评分关系 r_1 和信任关系 r_2 ,构建的线性回归模型具体如下:

$$y=a_0+sf_1x_1+sf_2x_2$$

其中, a_0,sf_1,sf_2 是回归系数。使用最小二乘法进行参数估计,建立如下方程组:

$$\begin{cases} \sum y=sf_1\sum x_1+sf_2\sum x_2 \\ \sum x_1y=sf_1\sum x_1^2+sf_2\sum x_1x_2 \\ \sum x_2y=sf_1\sum x_1x_2+sf_2\sum x_2^2 \end{cases}$$

对方程组求解得到参数估值的平均值为 $sf_1=2.0,sf_2=2.8$ 。

因此,对实验中构建的多关系社交网络进行社团结构划分时,设两种关系的关系比例系数为 $SF=(2.0,2.8)$,关系强度均为 1,设 K-means 聚类算法中聚类个数 $k=5$ 。社团结构的划分结果如图 7 所示。

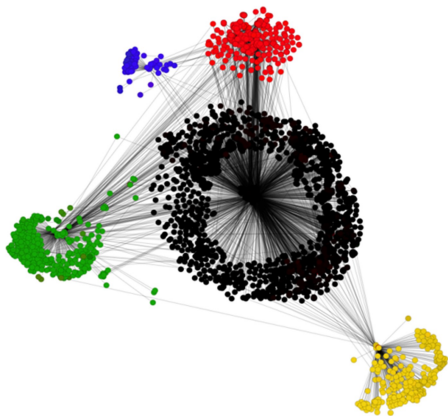


图 7 社团结构划分结果

Fig. 7 Result of community detection

6.2 实验结果分析对比

本文选取平均绝对误差(Mean Absolute Error,MAE)和均方根误差(Root Mean Squared Error, RMSE) 作为评价指

标来评价推荐结果的准确度。这两个指标是通过比对真实评分和预测评分之间的误差来评测推荐算法的准确度,其值越小说明推荐的效果越好。平均绝对误差(MAE) 和均方根误差(RMSE)的具体定义如下:

$$MAE=\frac{1}{N}|r_{i,j}-r'_{i,j}|$$

$$RMSE=\sqrt{\frac{1}{N}\sum(r_{i,j}-r'_{i,j})^2}$$

其中, $r_{i,j}$ 表示用户 i 对商品 j 的实际评分, $r'_{i,j}$ 表示用户 i 对商品 j 的预测评分, N 表示测试集中的评分数量。

最近邻 k 的选择会对实验结果产生影响,最近邻集合过小或者过大都会增加预测的偶然性,在实验过程中对本文所提算法分别设置最近邻 k 值为 10 和 30,结果如表 2 所列。可以看出, k 值为 30 时,MAE 和 RMSE 明显低于 k 值为 10 的情况,因此在后续各类推荐算法比较实验中,设置最近邻个数 $k=30$ 。

表 2 不同近邻数对算法的影响

Table 2 Influence of different neighbor numbers on algorithm

近邻数	MAE	RMSE
10	0.62	0.82
30	0.44	0.59

实验中,将本文所提算法(CDRec)与 3 种常用的社会化推荐算法 TDSRec^[18],SocialMF^[19]以及 TDRec^[20]在 Epinions 数据集和 FilmTrust 数据集上分别进行对比分析,结果如图 8 和图 9 所示。

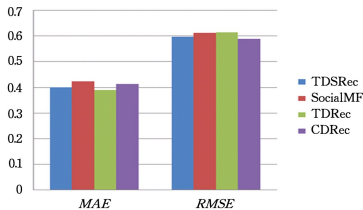


图 8 4 种推荐算法在 Epinions 数据集上的结果对比

Fig. 8 Result comparison of four recommendation algorithms on

Epinions dataset

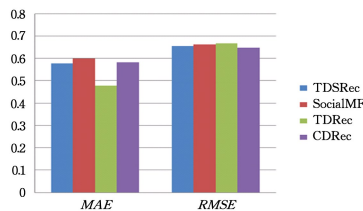


图 9 4 种推荐算法在 FilmTrust 数据集上的结果对比

Fig. 9 Results comparison of four recommendation algorithms on

FilmTrust dataset

由图 8 和图 9 可知,本文所提算法与其他社会化推荐算法在 MAE(平均绝对误差)上基本接近,在评价指标 RMSE(均方根误差)上略优于其他算法,这说明本文所提算法能够较为准确地对待推荐的商品进行预测评分。

对于不同规模的数据集,4 种社会化推荐算法的运行时间如图 10 所示。

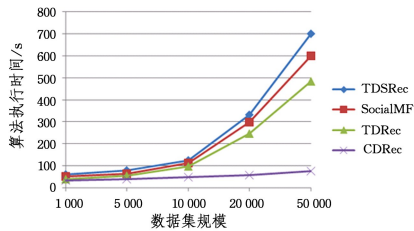


图 10 4 种推荐算法运行时间对比

Fig. 10 Execution comparison of four recommendation algorithms

由图 10 可知,随着数据集规模的增大,与其他 3 种社会化推荐算法相比,本文所提算法的执行时间明显较短;对于海量数据集,本文所提推荐算法的执行时间将缩短 90%左右。

综上所述,本文提出的算法在处理大数据样本集时,能够在保证推荐准确率的基础上提高推荐效率。

结束语 本文基于多子网复合复杂网络模型构建多关系社交网络,提出了一种融合多关系社交网络的协同过滤算法。该算法利用社团结构划分可以将多关系社交网络中相似度接近的用户划分到同一个社团中,并利用社团内部的用户构建用户-商品评分矩阵,进而极大地减小推荐算法的计算量。通过在真实数据集上的实验表明,该算法能够在保证不降低推荐准确率的基础上,提高推荐算法的效率。下一步将尝试在本文算法的基础上,在推荐系统中融入更多的社交关系,从而进一步提高推荐的准确率。

参 考 文 献

[1] LIU H X. A Survey of Collaborative Filtering Technique in Recommendation System[J]. Information Security and Technology, 2016, 7(3): 24-26.

[2] DONG C L, KE X S. Study on Collaborative Filtering Algorithm Based on User Interest Change and Comment [J]. Computer Science, 2018, 45(3): 213-217.

[3] ZHANG Y, LI S Y, TANG Y. Research on Heterogeneous Knowledge Fusion Algorithm under Big Data Environment[J]. Computer Technology and Development, 2017, 27(9): 12-16.

[4] MARGARIS D, VASSILAKIS C, GEORGIADIS P. Query personalization using social network information and collaborative filtering techniques[J]. Future Generation Computer Systems, 2018, 78(1): 440-450.

[5] BIN S. Research of Multi-relationship Online Social Networks Based on Multi-subnet Composited Complex Network Model [D]. Qingdao: Shandong University of Science and Technology, 2014.

[6] ZHAO Q Q. SPCF: A Memory Based Collaborative Filtering Algorithm via Propagation [J]. Chinese Journal of Computers, 2013, 36(3): 671-676.

[7] JIANG S, QIAN X, SHEN J. Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations [J]. IEEE Transactions on Multimedia, 2015, 17(6): 907-918.

[8] ROH T H, OH K J, HAN I. The collaborative filtering recommendation based on SOM cluster-indexing CBR [J]. Expert Systems with Applications, 2003, 25(3): 413-423.

[9] ZHANG H Y, DING F, JIANG L H. A Collaborative Filtering

Recommendation Method Based on Fuzzy Clustering [J]. Computer Simulation, 2005, 22(8): 144-147.

[10] SANTHINI M, BALAMURUGAN M, GOVINDARAJ M. Collaborative Filtering Approach for Big Data Applications based on Clustering [J]. International Journal of Mathematics Computer Science and Information Technology, 2015, 2(1): 202-208.

[11] WU M X, DONG L S, JIE Z Y. Research on Social Recommender Systems [J]. Journal of Software, 2015, 26(6): 1356-1372.

[12] CHANG W L, DIAZ A N, HUNG P C K. Estimating trust value: A social network perspective [J]. Information Systems Frontiers, 2015, 17(6): 1381-1400.

[13] PAN J C, ZHANG X M, WANG X. Improved Singular Value Decomposition Recommender Algorithm Based on User Reliability [J]. Journal of Chinese Computer Systems, 2016, 37(10): 2171-2176.

[14] SUN G X, BIN S. A New Opinion Leaders Detecting Algorithm in Multi-relationship Online Social Networks [J]. Multimedia Tools & Applications, 2017, 77(3): 1-13.

[15] JIANG M M, SUN G X, BIN S. Community Detection Algorithm in Multiple Relationships Online Social Network [J]. Journal of Frontiers of Computer Science and Technology, 2018, 7: 1-13.

[16] WANG Q, LI W, ZHANG X. Academic Paper Recommendation Based on Community Detection in Citation-Collaboration Networks [J]. Lecture Notes in Computer Science, 2016, 9932: 124-136.

[17] ZHANG H X, ZHEN L, ZHANG C T. An Improved Collaborative Filtering Recommendation Algorithm with Weighted Heterogeneous Information [J]. Journal of University of Electronic Science & Technology of China, 2018, 47(1): 112-116.

[18] GUO N N, WANG B L, HOU Y H. Collaborative Filtering Recommendation Algorithm Based on Characteristics of Social Network [J]. Journal of Frontiers of Computer Science and Technology, 2018, 2: 208-217.

[19] FORSATI R, MAHDAVI M, SHAMSFARD M. Matrix Factorization with Explicit Trust and Distrust Side Information for Improved Social Recommendation [J]. Acm Transactions on Information Systems, 2014, 32(4): 17.

[20] PARK C, KIM D, OH J. Improving top-K recommendation with truster and trustee relationship in user trust network [J]. Information Sciences, 2016, 374(3): 100-114.

[21] GUO L J, LIA N, ZHAO X W. Collaborative filtering recommendation algorithm incorporating social network information [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(3): 281-288.

[22] WANG M, MA J. A novel recommendation approach based on users' weighted trust relations and the rating similarities [J]. Soft Computing, 2016, 20(10): 3981-3990.

[23] LAUW H W, LIM E P, WANG K. Quality and Leniency in Online Collaborative Rating Systems [J]. Acm Transactions on the Web, 2012, 6(1): 1-27.

[24] LI W J, QI J, YU Z T. Social Recommendation Algorithm Integrating Trust Propagation and Singular Value Decomposition [J]. Computer Engineering, 2017, 43(8): 236-242.