第 45 卷 第 5 期 计 算 机 工 程 2019 年 5 月
Vol. 45 No. 5 Computer Engineering May 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)05-0175-07 文献标志码：A 中图分类号：TP399

基于多重图排序的用户冷启动推荐方法

毛明松，张富国

（江西财经大学 信息管理学院，南昌 330013）

摘 要：为使用户－物品评分、社会网络和社会化标签等异构信息融合到协同过滤推荐方法的最近邻寻找过程中，弥补冷启动用户单一维度信息的不足，提出一种多重图排序的冷启动推荐方法。通过分析用户之间可能存在的信任度构建关系网络，利用多重图排序模型得到目标用户的最近邻集合，进而产生目标用户的推荐列表。实验结果表明，与基于用户的协同过滤推荐方法相比，该方法能有效地提高冷启动用户的个性化推荐准确性和推荐覆盖率。

关键词：推荐方法；异构信息；冷启动；多重图模型；多元关系网络

中文引用格式：毛明松，张富国. 基于多重图排序的用户冷启动推荐方法[J]. 计算机工程，2019，45(5)：175-181.

英文引用格式：MAO Mingsong，ZHANG Fuguo. Recommendation method for user cold-start based on multi-graph rank[J]. Computer Engineering，2019，45(5)：175-181.

Recommendation Method for User Cold-start Based on Multi-graph Rank

MAO Mingsong，ZHANG Fuguo

（School of Information Management，Jiangxi University of Finance and Economics，Nanchang 330013，China）

【Abstract】In order to integrate the heterogeneous information such as user-item scoring，social network and social labeling into the nearest neighbor search process of collaborative filtering recommendation method，and to make up for the lack of information on the single dimension of cold-start users a cold-start recommendation method for multi-graph rank is proposed. A relational network is built by analyzing the trustworthiness that may exist between users，and the multi-graph sorting model is used to obtain the nearest neighbor set of the target user，thereby generating a recommendation list of the target user. Experimental results show that compared with the user-based collaborative filtering recommendation method，this method can effectively improve the personalized recommendation accuracy and recommended coverage rate of cold-start users.

【Key words】recommendation method；heterogeneous information；cold start；multi-graph model；multi-relational network

DOI：10.19678/j.issn.1000-3428.0050495

0 概述

随着互联网技术的飞速发展和现代电子商务的广泛应用，如何从海量数据中准确地获得用户感兴趣的信息是目前面临的一个难题。其中，推荐系统是缓解信息过载问题的有效工具之一，广泛应用于电子商务、视频分享、音乐电台、在线社交、个性化阅读和个性化广告等领域[1]。

根据信息过滤方法的不同，传统的推荐方法[2]主要有协同过滤推荐[3]、基于内容的推荐和混合推荐 3 种类型。协同过滤是应用与研究较多的推荐方法，但存在冷启动[4]等一系列问题，具体包括新用户问题和新项目问题。当新用户加入系统时，没有选择或只选择少量商品，系统无法给其提供准确的推荐服务；当新的商品加入系统时，由于协同过滤系统完全靠用户已有的选择信息进行推荐，因此只有该商品被用户打分，系统才可能推荐它。在用户偏好信息不足的条件下，把社会信任网络和社会化标签等多元异构信息融合到推荐系统决策过程中，有助于弥补用户-商品评分信息的不足，一定程度缓解了推荐系统的数据稀疏性和冷启动等问题。

本文提出一种基于多重图排序的冷启动推荐算法（Multi-graph Rank Recommendation Algorithm，MGRank）。该算法在考虑各网络的结构差异性基础上，利用多重图排序模型聚合用户的多重关系网络，得到新用户的最近邻集合，以产生目标用户的推荐列表。

基金项目：国家自然科学基金（71361012，71764006）；江西省自然科学基金（20161BAB201029）；江西省教育厅科技项目（GJJ170344，GJJ170319）。

作者简介：毛明松（1987—），男，讲师、博士，主研方向为个性化推荐、数据挖掘；张富国，副教授、博士。

收稿日期：2018-02-15 **修回日期：**2018-04-13 **E-mail：**redbird_mail@163.com

1 相关工作

近年来,国内外学者致力于研究如何缓解新用户的冷启动问题,对于没有评分或者只有少量评分的新注册用户,个性化推荐的应对方法按照使用数据源的差异分为以下两类:

1)充分利用现有新用户的少量物品评分,改进传统相似度计算方法。例如:文献[5]利用启发式算法度量用户之间的兴趣相似性,解决冷启动问题;文献[6]在用户和商品构成的二部图网络中添加虚拟连边,该方法改进了基于网络结构推荐算法在冷启动推荐上的性能。

2)向推荐系统提供额外信息。这些额外信息包括人口统计信息(性别、年龄和职业等)[7-8]、用户在注册时候提供的一些偏好调查信息[9-11]、社会化标签[12-14]和社交网络信息[15-16]等。如何使用额外的数据或者将额外的数据与原有数据相结合是研究的重点。文献[7]认为具有相同背景和性格的人更可能具有相似的偏好,利用人口统计学信息对用户进行分组,在分组中计算用户在不同属性上的相似度,再根据整体相似度取得目标用户的邻居集,然后将相似度加入到预测评分进行推荐。文献[11]在社会化标签系统中构建用户-物品-标签三部图,再通过物质扩散动力学方法来解决冷启动问题。文献[13]提出信任模型,该模型通过融合每个用户的信任度和不信任度来获得可信任用户(专家),然后由专家向新用户推荐物品。文献[15]通过计算用户之间的信任度来代替相似度计算,取得相似友邻集,实现信息推荐。文献[17]提出冷启动方法,通过社区检测算法在社交网络上区分用户的相似度。

上述方法在一定程度上缓解了新用户的冷启动问题,但对于第 1 种方法,新用户评分较少,推荐性能难以改善;对于第 2 种方法,存在单一数据源信息过少的问题。例如:在注册时,人口统计信息或对部分物品偏好信息的收集需要用户的积极配合,但这种调查通常会使用户非常厌烦而不愿提供,因此文献[18]研究如何选择较少的问题来捕捉用户的兴趣爱好。对冷启动用户,虽然单个数据源的信息不多,但通过融合多元异构信息,很可能会成为发掘用户兴趣偏好的一种有效途径。

2 多重图排序推荐模型

2.1 多重图构建

在多元异构信息环境中的用户之间存在着多种关系网络,每一种关系可以构成一个简单图,为了区分多重图结构和一般联合图,下文给出其定义。

定义 1(单一关系网络) 在一个推荐系统中,单一关系网络定义为一个简单图 $G=\{U,E\}$,其中,顶点集合 U 为系统全体用户,边集合 $E\subseteq U\times U$ 为用户之间的某一种关联关系,其中 2 个相邻用户之间存在有向关系,记作 $[u_i,u_j]\in E$。

用户之间的关系网络可能呈现多种形式,比如有权重或无权重、有向或无向等。本文将无权重的社交网络视作权重均为 1 的有权网络,将无向关系均视作 2 条有向关系,相应地,为每一个加权网络定义一个大小为 $|U|\times|U|$ 的权重矩阵 $\boldsymbol{W}$,其中如果从用户 u_i 到用户 u_j 存在一个有向连接,则有 $w_{ij}=([u_i,u_j])$;否则 $w_{ij}=0$。

为融合用户之间的不同关系网络,一种简单方法是通过合并用户之间的每一种关系构建一个联合图[19]。

定义 2(联合图模型) 给定若干子图 $G_1=\{U,E_1\},G_2=\{U,E_2\},\cdots,G_z=\{U,E_z\}$ 为同一用户集 U 的 Z 个关系网络,其中,$E_1,E_2,\cdots,E_Z\subseteq U\times U$ 为各图的边集合,每一个子图 $G_k(k=1,2,\cdots,z)$ 对应一个权重矩阵 $\boldsymbol{W}_k$,一个联合图 $G'=\{U,E'\}$ 是以用户集 U 为顶点的一个新的简单图,E' 为原来各个子图边的并集,即 $E'=\bigcup_{k=1}^{Z}E_k$。相应地,联合图的边权重由各个子图的权重按照特定方法进行聚合,即:

$$w'(e)=g(w_1(e),w_2(e),\cdots,w_Z(e)),\forall e\in E' \tag{1}$$

其中,g 为聚合函数。联合图模型的缺点在于只是简单地将多个子网络进行加权合并,无法保留各个子图原有的结构信息,但多重图模型能够避免该缺点[20]。

定义 3(多重图模型) 一个多重图 $G=\{U,E\}$ 是在用户集 U 上的一个特殊的图结构,其边为全体子图边的多重集合,即有 $E=\biguplus_{k=1}^{Z}E_k$。

多重图模型可以视作将各个子图重叠起来而非合并成一个新的简单图结构。本文用下标 i 和 j 区分不同用户(顶点),用下标 k 和 l 区分不同子图。

2.2 网络间的结构差异

网络与网络之间的结构差异性或相似性已经在相关领域中获得不少研究。本文引入文献[18-19]提出的平均邻居相似度(Average Similarity of Neighbours,ASN)作为比较不同网络结构的指标。由于 ASN 指标不适用于加权网络,因此本文对其改进,给出如下定义:

$$ASN(A,B)=\frac{\sum_i DEG_{AB}(i)}{\sum_i DEG_A(i)+DEG_B(i)} \tag{2}$$

其中,$DEG_A(i)$ 和 $DEG_B(i)$ 分别为在图 A 和图 B 中节点 i 链边的权重之和,$DEG_{AB}(i)$ 为节点 i 在图 A 和图 B 中对其在 2 个图中的共同邻居链边的权重之和。显然,可以得到 $DEG_{AB}(i)\leqslant DEG_A(i)+DEG_B(i)$,且等号成立的充分必要条件是该节点在图 A 和图 B 里的邻居节点完全相同。本文定义网络 A 和 B 之间的结构差异性 $\delta(A,B)$ 为:

$$\delta(A,B)=1-ASN(A,B) \tag{3}$$

如果发现 2 种关系网络之间的结构差异性较小(比如 $\delta<0.1$),即 2 个网络的大部分边都重叠,则只取其中一种网络作为输入信息。具体原因如下:

1)聚合具有相同结构的用户关系网络不能消除数据稀疏性问题。

2)具有相同结构的关系网络可能来源于高度相关的数据,将它们结合可能只是对同一输入信息的重复利用,导致最终推荐结果的不正确。

同一顶点集的 2 个网络之间结构性差异大是指边的重叠部分少,信息较为互补,而不是指相同顶点之间边的权重不同。在推荐系统中,应注重对具有信息互补资源的融合,以消除单一资源下冷启动用户的数据稀疏性。

2.3　多重图排序模型

2.3.1　简单图排序

在图排序问题中,给定一个简单图 $G=(V,E)$,其中,$V=\{v_1,v_1,\cdots,v_n\}$ 为顶点集,$E\subseteq V\times V$ 为边集,$\boldsymbol{W}$ 为边的权重矩阵。给定一个初始查询向量 $\boldsymbol{y}\in\mathbb{R}^{|V|}$,第 i 个元素 y_i 为节点 v_i 的查询初始值。查询向量也可以理解为一个在顶点空间的排序函数 $\boldsymbol{y}:V\to\mathbb{R}$,使得 $y(v_i)=y_i$。图排序问题则可表述为寻找一个新的在顶点空间的排序函数 $\boldsymbol{f}:V\to\mathbb{R}$,使其足够平滑并且同时接近给定的初始排序函数 $\boldsymbol{y}$。同样地,排序函数 $\boldsymbol{f}$ 可对应一个排序向量 $\boldsymbol{f}\in\mathbb{R}^{|V|}$,图排序问题可以形式化为最小化损失函数[21-23],表示为:

$$\min_{\boldsymbol{f}:V\to\mathbb{R}} Q(\boldsymbol{f})=S(\boldsymbol{f})+\mu\hat{R}(\boldsymbol{f};\boldsymbol{y}) \tag{4}$$

式(4)等号右边由两部分组成,第一部分 $S(\boldsymbol{f})$ 为排序函数 $\boldsymbol{f}$ 的平滑度,代表排序函数的结构损失项;第二部分 $\hat{R}(\boldsymbol{f};\boldsymbol{y})$ 为排序函数 $\boldsymbol{f}$ 与初始查询函数 $\boldsymbol{y}$ 的排序误差,代表排序函数的经验损失项,并引入参数 $\mu>0$,以平衡这 2 个损失项。

排序误差项 $\hat{R}(\boldsymbol{f};\boldsymbol{y})$ 通常设为 $\boldsymbol{f}$ 与 $\boldsymbol{y}$ 的 l_2 范式误差[21],其矩阵形式如下:

$$\hat{R}(\boldsymbol{f};\boldsymbol{y})=\|\boldsymbol{f}-\boldsymbol{y}\|^2=(\boldsymbol{f}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f}-\boldsymbol{y}) \tag{5}$$

排序函数应该足够平滑,即对于 2 个“关系非常密切”的顶点,其排序结果应相差不大。引入权重矩阵 $\boldsymbol{W}$ 和定点的度矩阵 $\boldsymbol{D}$,根据文献[21],平滑函数 $\boldsymbol{S}(\boldsymbol{f})$ 计算公式为:

$$\boldsymbol{S}(\boldsymbol{f})=\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{f} \tag{6}$$

其中,矩阵 $\boldsymbol{A}$ 定义为 $\boldsymbol{A}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。

令损失函数 $\boldsymbol{Q}(\boldsymbol{f})$ 偏导为 0,可得:

$$\left.\frac{\partial Q}{\partial \boldsymbol{f}}\right|_{\boldsymbol{f}=\boldsymbol{f}^*}=(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{f}^*+\mu(\boldsymbol{f}^*-\boldsymbol{y})=0 \tag{7}$$

定义参数 $\alpha=1/(\mu+1)\in(0,1)$,由式(7)可得最优排序结果 $\boldsymbol{f}^*$ 为:

$$\boldsymbol{f}^*=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{A})^{-1}\boldsymbol{y}\propto(\boldsymbol{I}-\alpha\boldsymbol{A})^{-1}\boldsymbol{y} \tag{8}$$

由于正常数项 $1-\alpha$ 不影响 $\boldsymbol{f}^*$ 的排序顺序,因此式(8)约简掉该项。

2.3.2　多重图排序模型

多重图模型的排序和简单图排序问题的整体目标一致,即寻找最优排序函数 $\boldsymbol{f}:V\to\mathbb{R}$,使其尽可能平滑并与初始查询函数 y 保持一致,因此多重图排序仍然可化为最小化损失函数问题。对于多重图模型,排序误差项 $\hat{R}(\boldsymbol{f};\boldsymbol{y})$ 为式(5),但平滑函数 $S(\boldsymbol{f})$ 需重新定义。

在一个由 Z 个简单子图构建的多重图模型中,一对节点之间可以由多条边连接。因此,2 个节点 u_i 和 u_j 之间的多重关系可以表述为:

$$[u_i,u_j]\leftarrow\begin{pmatrix}w_{ij}^1\\w_{ij}^2\\\vdots\\w_{ij}^Z\end{pmatrix} \tag{9}$$

其中,w_{ij}^k 为 $w_k([u_i,u_j])$ 的简化表达,表示该对节点在第 k 个子图中边的权重。

本节首先初步估计一对节点之间的整体关系,即引入一个虚拟边函数,将一对节点之间的关系映射为一个实数[19]。多重图节点之间同时存在多种关系,而不同关系网络之间可能具有不同的结构差异性,因此边函数既考虑用户节点间的“网络内关系”,也考虑各个子图之间的“网络间关系”。基于上述分析,引入网络结构差异性测度,给出多重图模型的边函数 $\pi:U\times U\to\mathbb{R}$ 的定义如下:

$$\pi([u_i,u_j];\boldsymbol{\Delta})=\frac{1}{2}\sum_{k,l}^{Z}w_{ij}^k w_{ij}^l\Delta_{kl} \tag{10}$$

其中,矩阵 $\boldsymbol{\Delta}$ 是大小为 $Z\times Z$ 的子图间的多样性矩阵,定义为:

$$\Delta_{kl}=\begin{cases}\delta(G_k,G_l), & k\neq l\\1, & k=l\end{cases} \tag{11}$$

在式(10)中,当 $k\neq l$ 时,Δ_{kl} 为子图 G_k 和子图 G_l 的结构差异性;当 $k=l$ 时,Δ_{kl} 取值为 1。基于式(10),可得到一个虚拟边矩阵 $\boldsymbol{\Pi}$,其每个元素为 $\boldsymbol{\Pi}_{ij}=\pi([u_i,u_j])$,计算公式为:

$$\boldsymbol{\Pi}=\frac{1}{2}\sum_{k,l=1}^{Z}\Delta_{kl}\boldsymbol{W}_k\boldsymbol{W}_l \tag{12}$$

根据虚拟边矩阵 $\boldsymbol{\Pi}$,定义节点输出度 $d^+(u_i)$ 和输入度 $d^-(u_i)$ 分别为:

$$d^+(u_i)=\sum_{j\neq i}\pi([u_i,u_j]) \tag{13}$$

$$d^-(u_i)=\sum_{j\neq i}\pi([u_j,u_i]) \tag{14}$$

此外,给出 2 个相应的对角矩阵 $\boldsymbol{D}_+$ 和 $\boldsymbol{D}_-$,其对角元素分别为 $(D_+)_{ii}=d^+(u_i)$ 和 $(D_-)_{ii}=d^-(u_i)$。

参考文献[22]简单图排序的正规化思想,排序平滑损失函数 $S(f)$ 由排序函数在每一条边上的偏导

数构成。对于多重图模型,排序函数在某一顶点处虚拟边的偏导为:

$$\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_i}=\sqrt{\frac{\pi([u_i,u_j])}{d^+(u_i)}}f(u_i)-\sqrt{\frac{\pi([u_i,u_j])}{d^-(u_j)}}f(u_j) \tag{15}$$

排序函数在某一顶点处的一条接收的虚拟边的偏导为:

$$\left.\frac{\partial f}{\partial[u_j,u_i]}\right|_{u_i}=\sqrt{\frac{\pi([u_j,u_i])}{d^-(u_i)}}f(u_i)-\sqrt{\frac{\pi([u_j,u_i])}{d^+(u_j)}}f(u_j) \tag{16}$$

由此可知,$\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_j}=-\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_i}$。排序函数在某一顶点处的局部扰动为:

$$\|\nabla_{u_i}f\|=\sqrt{\frac{1}{2}\left[\sum_{j\neq i}\left(\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_i}\right)^2+\sum_{j\neq i}\left(\left.\frac{\partial f}{\partial[u_j,u_i]}\right|_{u_i}\right)^2\right]} \tag{17}$$

排序函数 f 的平滑度函数则定义为其在所有顶点处的局部扰动之和。去掉重复的边,平滑度函数定义为:

$$S(f)=\frac{1}{2}\sum_{u\in U}\|\nabla_u f\|^2 \tag{18}$$

展开可得:

$$\begin{aligned}S(f)&=\frac{1}{2}\sum_{i=1}^{n}\|\nabla_{u_i}f\|^2=\\&\frac{1}{4}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_i}\right)^2+\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left.\frac{\partial f}{\partial[u_j,u_i]}\right|_{u_i}\right)^2\right]=\\&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left.\frac{\partial f}{\partial[u_i,u_j]}\right|_{u_i}\right)^2=\\&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\pi_{ij}}{d_i^+}f_i^2+\frac{\pi_{ij}}{d_j^-}f_j^2-2\frac{\pi_{ij}f_if_j}{\sqrt{d_i^+d_j^-}}\right)=\\&\frac{1}{2}\sum_{i=1}^{n}f_i^2+\frac{1}{2}\sum_{j=1}^{n}f_j^2-\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}f_if_j}{\sqrt{d_i^+d_j^-}}=\\&\boldsymbol{f}^{\mathrm{T}}\boldsymbol{f}-\boldsymbol{f}^{\mathrm{T}}D_+^{-1/2}\boldsymbol{\Pi}D_-^{-1/2}\boldsymbol{f}\triangleq\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\end{aligned} \tag{19}$$

其中,矩阵 $\boldsymbol{S}$ 定义为:

$$\boldsymbol{S}=D_+^{-1/2}\boldsymbol{\Pi}D_-^{-1/2} \tag{20}$$

因此,多重图排序问题转化为以下损失函数:

$$\min_{f:U\to\mathbb{R}}Q(f)=\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}+\mu(\boldsymbol{f}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f}-\boldsymbol{y}) \tag{21}$$

多重图排序的损失函数与简单图优化的损失函数形式相同,采取相同的求解过程可得多重图排序的最优排序结果为:

$$\boldsymbol{f}_m=(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}\boldsymbol{y} \tag{22}$$

其中,参数 α 取值范围为(0,1),初始查询向量 $\boldsymbol{y}$ 为各个节点的初始排序得分,用来对当前目标节点与其他节点的初始联系程度进行初步判断。对于简单图,查询向量 $\boldsymbol{y}$ 可由权重 $\boldsymbol{W}$ 给出,代表目标节点与其他节点的直接联系。本文利用边函数构建多重图模型,可表示为:

$$y_i=\begin{cases}\frac{1}{2}Z^2, & u_i=u_a\\ \pi([u_a,u_i]), & u_i\neq u_a\end{cases} \tag{23}$$

根据式(10),多重图模型的边函数满足 $\pi([u_a,u_i])\leqslant\frac{1}{2}Z^2$,因此式(23)能够确保目标用户 u_a 所对应的初始查询值最大。

3 冷启动推荐算法

本文算法的输入信息是推荐系统中用户之间的各种关系,可以是显性的用户社交关系,也可以是隐性的各种联系,输出信息是目标用户 u_a 的推荐列表,其具体步骤如下:

步骤 1 利用不同数据源分析用户之间可能存在的相似性或者信任度来构建各类初始关系网络。例如:通过提取用户在线好友关系构建一个友邻网络[15];根据评分相似度构建用户相似度网络;由用户其他历史行为或记录构建更广泛的隐性关联,如通过标签信息提取用户之间的偏好相似度[11-12]。

步骤 2 构建多重图模型。多重图模型由收集的各类单一网络构建,但是初步收集的各种关系网络通常比较稀疏。为丰富初始信息,可以利用用户信任的传递性对非相邻用户之间的间接联系进行初步估计,对单一网络进行扩展。采用的方法通常包括网络搜索和随机行走等。本文采用最简单的两层广度优先搜索策略进行初始网络扩展,即对于非相邻节点 u 和 v,如果存在一个中间节点 m 将其相连,则将这条路径上的两条边的权重之积作为其间接联系的初步估计,可表达为 $\hat{w}(u,v)=\theta\cdot w(u,m)\cdot w(m,v)$,其中,$\theta$ 为信任传递的衰减系数[24]。如果存在多个中间节点,则取所有估计结果的平均值为最终结果。对于衰减稀疏 θ,一般取 0.5 ~ 0.8,本文取为 0.8。

将所有单一网络进行扩展之后,通过计算初始网络之间的结构差异性(式(3)),对初始网络进行筛选,构建一个包含多种联系的多重图模型。

步骤 3 训练多重图排序模型所需中间矩阵和参数。根据多重图排序模型的解(式(22)),训练得到矩阵 $\boldsymbol{S}$ 和参数 α 以及其他中间矩阵,作为输入信息预先存储并定时更新。

步骤 4　运用多重图排序方法找出目标用户在多重关系网络环境下的最近邻用户。通过式(22)求解可得到目标用户的最优排序向量 $\boldsymbol{f}$,并从中选取排序分最大的 K 个用户组成邻居用户,记为 $Neib(u_a)$。

步骤 5　根据协同过滤思想,通过式(24)预测目标用户 u_a 对潜在物品 i 的评分 $\hat{r}_{u_a,i}$,并产生推荐列表。

$$\hat{r}_{u_a,i}=\bar{r}_{u_a}+\frac{\sum_{u\in Neib(u_a)} f(u)(r_{u,i}-\bar{r}_u)}{\sum_{u\in Neib(u_a)} f(u)} \tag{24}$$

其中,$\bar{r}_u$ 为用户 u 的评分均值,$f(u)$ 为目标用户 u_a 的最优排序向量 $\boldsymbol{f}$ 对应的用户 u 的排序分。

4　实验结果与分析

4.1　数据集

本文实验数据集来自于全世界著名的社会音乐平台 Last. fm(http://www. last. fm),last. fm 是以英国为总部的网络电台和音乐社区,有遍布 232 个国家超过 15 000 000 的活跃听众,因为同时包含用户对物品的偏好数据、用户之间的社交关系以及用户所做的标签数据,被广泛用于异构信息环境下的推荐方法性能测试。Last. fm 数据集由第二届推荐系统领域信息异构和融合专题讨论会公开发布,包括了1 892 个用户、17 632 个艺术家和 186 479 个标签。该数据集没有用户对物品(艺术家)的直接评分值但是可以将用户对艺术家音乐的收听计数作为其隐性评分值。本文实验将每个用户收听的音乐按照其收听次数从多到少排序,排在前 1/5 的音乐视作评分值为 5,接下来 1/5 视作评分值为 4,依次类推至最后 1/5 视作评分值为 1。在 Last. fm 数据集中,用户可以互相加为好友,从而形成一种在线朋友关系。此外,该数据集还提供用户对艺术家标记的标签信息。数据集中基本信息如表 1 所示。

表 1　Last. fm 数据集基本信息

数据	使用数	稀疏度/%
用户数	1 892	—
物品数	17 632	—
标签数	11 946	—
打分记录	92 834	99.72
朋友关系	25 424	99.29
用户打标签记录	186 479	99.17

实验中随机选取 20% 用户为冷启动用户,这些用户的收听记录、朋友关系和标签个数超过 5 个用户被移到测试集,剩下用户和非冷启动用户的所有数据构成训练集。

4.2　评价标准

本文选取准确度和覆盖率作为评价标准,具体分析如下:

1)预测准确度

推荐方法是否能够准确预测用户对物品的打分通常由预测值与真实评分值之间的误差来衡量。常用的评价指标有平均绝对误差(Mean Absolute Error,MAE)[25] 和均方根误差(Root Mean Squared Error,RMSE)[26],后者对较大的绝对误差惩罚较重。本文选取 RMSE 作为准确度的评价指标,定义为:

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{|E^P|}\sum_{(u,a)\in E^P}(\hat{r}_{ua}-r_{ua})^2} \tag{25}$$

其中,$\hat{r}_{ua}$ 表示用户 u 对物品 a 的预测评分,r_{ua} 表示用户 u 对物品 a 的真实评分,E^P 表示测试集。

2)覆盖率

覆盖率指标[4] 是推荐系统向用户推荐的物品能够覆盖全部物品的比例。如果一个推荐系统的覆盖率较低,则该推荐系统很有可能会因为它推荐的范围比较小而导致用户的满意度降低。推荐系统的覆盖率低说明用户可以选择的物品相对较少,比较局限。

覆盖率[25] 包括预测覆盖率、推荐覆盖率和种类覆盖率 3 种。本文采用预测覆盖率,其表示系统可以预测评分的商品占所有商品的比例,定义为:

$$COV_p=\frac{N_d}{N} \tag{26}$$

其中,N_d 表示推荐系统能够预测评分的不同物品的数目,N 表示所有物品的数目。

4.3　结果分析

本文选取以下 4 种推荐算法作为对比算法:

1)基于用户的协同过滤推荐算法[27],简称 UCF,实验采用 Cosine 相似度来计算用户之间的相似度。

2)基于信任关系的推荐算法,简称 sTrust,实验参照文献[28]给出的基于信任关系的推荐算法。

3)基于标签信息的推荐算法,简称 sTag。实验参照文献[29]利用用户标签信息构建用户之间的基于标签的相似度关系。

4)多重图排序推荐算法,简称 MGRank。

实验中的推荐近邻数 k 的变化值分别为 5、10、20、30 和 50。本文 MGRank 模型中参数 α 需要通过验证性实验进行优化设定。在所选数据集上测试 α 的不同取值(α 为 0.1 ~ 1,间隔为 0.01)对 MGrank 方法的预测准确度 RMSE 的影响。

图 1 为推荐友邻数 $k=50$ 下模型参数 α 对 MGRank 算法 RMSE 的影响。随着参数 α 的增大,本文算法的预测误差呈现不断下降的趋势并在 0.97 时达到最低。根据损失函数式(4)可知此时排序结构损失项(排序平滑度)比较重要,而经验损失项(与初始查询向量一致性)则不太重要。

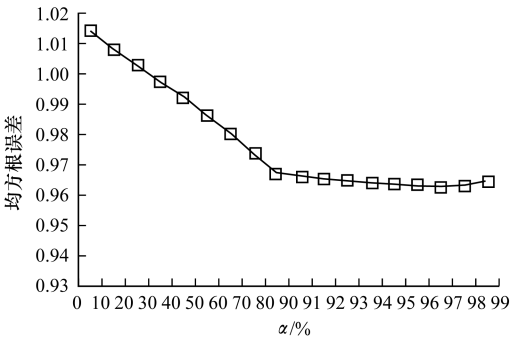


图 1 MGRank 算法在不同参数 α 下的均方根误差

不同推荐友邻数 4 种推荐算法的 RMSE 对比结果如图 2 所示,本文算法比 UCF 算法、sTrust 算法和 sTag 算法的 RMSE 值要小,表明融合多种资源的多重图推荐模型能够有效提高预测准确度。图 3所示为各推荐算法的推荐覆盖率,可以看出,MGRank 算法显著提高了冷启动用户的推荐覆盖率。因此,本文算法不仅能够提高推荐覆盖率,也能提高推荐准确度。

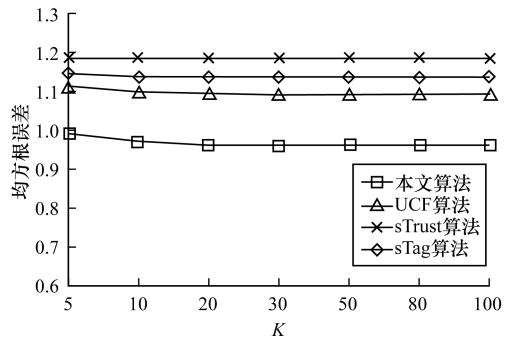


图 2 各推荐算法均方根误差对比结果

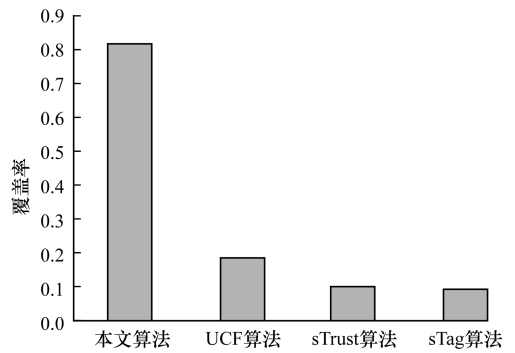


图 3 各推荐算法推荐覆盖率结果

5 结束语

本文针对推荐系统中用户社会网络多样化的特性提出多重图排序方法。通过网络间结构差异性指标分析网络间的多样性,利用用户之间的网络关系和网络间多样性,给出多图排序算法,以帮助用户从多重网络中找到最近邻居用户。在此基础上,采用协同过滤方法找出目标用户可能喜欢的物品。实验结果验证了该方法的有效性。但本文方法不能有效处理在线系统的多元关系,如用户-物品-标签。因此,下一步将引入超图结构和超图排序算法,解决个性化推荐系统中的冷启动问题。

参考文献

[1] LU Jie, WU Dianshuang, MAO Mingsong, et al. Recommender system application developments[J]. Decision Support Systems, 2015, 74(C): 12-32.

[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] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 1999: 230-237.

[4] SCHEIN A I, POPESCUL A, UNGAR L H, et al. Methods and metrics for cold-start recommendations[C]// Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2002: 253-260.

[5] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1): 37-51.

[6] ZHANG Fuguo, ZENG An. Improving information filtering via network manipulation[EB/OL]. [2018-01-08]. https://128.84.21.199/pdf/1208.6119.pdf.

[7] LIKA B, KOLOMVATSOS K, HADJIEFTHYMIADES S. Facing the cold start problem in recommender systems[J]. Expert Systems with Applications, 2014, 41(4): 2065-2073.

[8] SAFOURY L, SALAH A. Exploiting user demographic attributes for solving cold-start problem in recommender system[J]. Lecture Notes on Software Engineering, 2013, 1(3): 303-307.

[9] POIRIER D, FESSANT F, TELLIER I. Reducing the cold-start problem in content recommendation through opinion classification[C]//Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Washington D. C., USA: IEEE Press, 2010: 204-207.

[10] ZHANG Xi, CHENG Jian, QIU Shuang, et al. DualDS: a dual discriminative rating elicitation framework for cold start recommendation[J]. Knowledge-Based Systems, 2015, 73: 161-172.

[11] ZHANG Zike, LIU Chuang, ZHANG Yicheng, et al. Solving the cold-start problem in recommender systems with social tags[EB/OL]. [2018-01-08]. http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/11/201 0112113532947455.pdf.

[12] JI Ke, SHEN Hong. Addressing cold-start: scalable recommendation with tags and keywords[J]. Knowledge-Based Systems, 2015, 83: 42-50.

[13] CHEN C C, WAN Yuhao, CHUNG M C, et al. An effective recommendation method for cold start new users using trust and distrust networks[J]. Information Sciences, 2013, 224: 19-36.

[14] GOLBECK J. Combining provenance with trust in social networks for semantic web content filtering [C]// Proceedings of International Provenance and Annotation Workshop. Berlin,Germany:Springer,2006:101-108.
[15] VAIRACHILAI S, KAVITHADEVI M K, RAJA M. Alleviating the cold start problem in recommender systems based on modularity maximization community detection algorithm [J]. Circuits and Systems, 2016, 7(8):1268-1279.
[16] MASSA P, BHATTACHARJEE B. Using trust in recommender systems: an experimental analysis [C]// Proceedings of International Conference on Trust Management. Berlin,Germany:Springer,2004:221-235.
[17] 宋波伟,伍伟丽.基于信任社交圈的好友推荐算法[J].计算机工程,2017,43(5):149-155.
[18] SUN Mingxuan, LI Fuxin, LEE J, et al. Learning multiple-question decision trees for cold-start recommendation [C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York,USA:ACM Press,2013:445-454.
[19] JACOB Y,DENOYER L,GALLINARI P. Classification and annotation in social corpora using multiple relations [C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York,USA: ACM Press,2011:1215-1220
[20] GJOKA M,BUTTS C T,KURANT M,et al. Multigraph sampling of online social networks[J]. IEEE Journal on Selected Areas in Communications,2011,29(9):1893-1905.
[21] AGARWAL S. Ranking on graph data[C]//Proceedings of the 23rd International Conference on Machine Learning. New York,USA:ACM Press,2006:25-32.
[22] ZHOU Dengyong, SCHÖLKOPF B. A regularization framework for learning from graph data [EB/OL]. [2018-01-08]. https://www.microsoft.com/en-us/research/wp-content/uplo ads/2017/01/RFLG.pdf.
[23] MAO Kuang, FAN Ju, SHOU Lidan, et al. Song recommendation for social singing community [C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York, USA: ACM Press, 2014: 127-136.
[24] SHAMBOUR Q,LU Jie. A trust-semantic fusion-based recommendation approach for E-business applications [J]. Decision Support Systems,2012.54(1):768-780.
[25] HERLOCKER J L,KONSTAN J A,TERVEEN L G,et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems,2004,22(1): 5-53.
[26] 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(2):163-175.
[27] RESNICK P,IACOVOU N,SUSHAK M,et al. Grouplens:an open architecture for collaborative filtering of netnews[C]// Proceedings of ACM Conference on Computer Supported Cooperative Work. New York, USA: ACM Press, 1994: 175-186.
[28] MASSA P, AVESANI P. Computing with social trust[M]. Berlin,Germany:Springer,2009,259-285.
[29] ZHEN Yi, LI Wujun, YEUNG D Y. TagiCoFi: tag informed collaborative filtering[C]//Proceedings of the 3rd ACM Conference on Recommender Systems. New York,USA:ACM Press,2009:69-76.

编辑　赵　辉

(上接第 174 页)

[8] ZIDA S,FOURNIER-VIGER P,LIN C W,et al. EFIM: a fast and memory efficient algorithm for high-utility itemset mining[J]. Knowledge and Information Systems, 2017, 51(2):595-625.
[9] LIN Chunwei,YANG Lu,FOURNIER-VIGER P,et al. A binary PSO approach to mine high-utility itemsets[J]. Soft Computing,2017,21(17):5103-5121.
[10] WU Mingtai, ZHAN Justin, LIN Chunwei. An ACO-based approach to mine high-utility itemsets [J]. Knowledge-Based Systems,2017,116:102-113.
[11] SONG We, HUANG Chaomin. Mining high utility itemsets using bio-inspired algorithms: a diverse optimal value framework [J]. IEEE Access, 2018, 6 (1): 19568-19582.
[12] WU Chengwei,SHIE B E,TSENG V S,et al. Mining top-K high utility itemsets[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2012:78-86.
[13] DAWAR S, SHARMA V, GOYAL V. Mining top-k high-utility itemsets from a data stream under sliding window model[J]. Applied Intelligence,2017,47(4): 1240-1255.
[14] 李慧,刘贵全,瞿春燕.频繁和高效用项集挖掘[J].计算机科学,2015,42(5):82-87.
[15] ZIHAYAT M,An A. Mining top-k high utility patterns over data streams [J]. Information Sciences, 2014, 285(1):138-161.
[16] RYANG H,YUN U. Top-k high utility pattern mining with effective threshold raising strategies [J]. Knowledge-Based Systems,2015,76:109-126.
[17] TSENG V S,WU C W,FOURNIER-VIGER P,et al. Efficient algorithms for mining top-k high utility itemsets[J]. IEEE Transactions on Knowledge and Data Engineering,2015,28(1):54-67.
[18] 吴倩,王林平,罗相洲,等.一种快速挖掘 top-k 高效用模式的算法[J].计算机应用研究,2017,34(11): 3303-3307.
[19] 王乐,冯林,王水.不产生候选项集的 top-k 高效用模式挖掘算法[J].计算机研究与发展,2015,52(2): 445-455.
[20] FOURNIER V P,LIN C W,GOMARIZ A,et al. The SPMF open-source data mining library version 2[C]// Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin,Germany:Springer,2016:36-40.

编辑　樊丽娜