文章编号:1007－6735(2019)02－0174－09 DOI: 10.13255/j.cnki.jusst.2019.02.012

基于加权二部图的个性化方案推荐

杨 珍， 耿秀丽
（上海理工大学 管理学院，上海 200093）

摘要：针对传统协同过滤算法难以解决数据稀疏性、冷启动及用户兴趣各异的问题，提出了基于加权二部图的个性化推荐方法，解决个性化设计方案推荐问题。采用加权二部图，基于用户特征和方案特征的评分，对用户和方案分类，减轻数据稀疏性，形成用户−方案规则库；采用加权网络的协同过滤算法，计算新用户特征与用户−方案规则库中用户特征的改进相似度，通过 Top-N 方法筛选高相似的方案集进行推荐，解决冷启动和用户兴趣各异的问题。最后与传统协同过滤算法、加权二部图个性化推荐进行比较，证明该方法的有效性和实用性。

关键词：加权二部图；加权网络；协同过滤算法；改进相似度；Top-N 方法
中图分类号：TH 122; N 94 **文献标志码：**A

Personalized Solution Based on Weighted Bipartite Graphs

YANG Zhen, GENG Xiuli
(*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: As the most widely used recommendation algorithm, the traditional collaborative filtering can hardly solve the problems of data sparsity, cold start and different user interests. Aiming at these three problems, a personalized recommendation method based on weighted bipartite graphs was proposed to solve the problem of recommendation of personalized design scheme. A weighted bipartite graph was used to classify users and schemes based on user characteristics and scoring features to reduce the data sparsity and form a user-scheme rule base. A collaborative filtering algorithm based on weighted networks was used to calculate the improved similarity of user features between new user characteristics and user-scheme rules in the library, and recommended by the Top-N method to screen high similar solution sets to solve the problems of cold starting and different user interests. Finally, compared with the traditional collaborative filtering algorithm and weighted bipartite graph, the validity and practicability of the method were proved.

Keywords: *weighted bipartite graph*; *weighted network*; *collaborative filtering algorithm*; *improved similarity*; *Top-N method*

收稿日期: 2017−12−07
基金项目: 国家自然科学基金资助项目（71301104，71271138）；高等学校博士学科点专项科研基金资助项目（20133120120002）；上海市教委科研创新项目（14YZ088）；上海高校一流学科建设计划（S1201YLXK）；沪江基金资助项目（A14006）
第一作者: 杨 珍（1994−），女，硕士研究生．研究方向：管理科学与工程等．E-mail：jenny_0513@163.com
通信作者: 耿秀丽（1984−），女，副教授．研究方向：数据挖掘、个性化推荐等．E-mail：xiuliforever@163.com

随着互联网的快速发展，数据日益以几亿级别的速度增长，导致所需信息的获取难度提升，推荐系统就是为了解决信息超载问题而出现的，它主要根据用户历史行为推荐相似产品或服务。电子商务类网站的推荐系统已经较为成熟，如阿里巴巴、亚马逊等，都采用推荐技术向用户推荐其可能感兴趣的产品，挖掘潜在客户，从而提高产品收益。方案设计是根据顾客需求确定设计方案的过程，随着市场环境的变化，客户需求越来越有个性化，企业需要设计个性化的产品/服务方案。传统产品/服务设计方案的生成方法主要有配置技术、规则技术等，本文采用电子商务领域成熟的推荐技术研究个性化方案推荐技术，用于提高方案设计的精度和效率。

推荐技术是推荐系统的核心技术，根据以往顾客需求偏好、行为信息等建立兴趣模型，并用推荐算法向新顾客推荐其未曾产生行为的产品或服务。协同过滤算法 (collaborative filtering algorithm, CFA) 作为推荐技术中应用比较成熟的算法之一，其基本思想是根据顾客需求行为的相似度来推荐资源。协同过滤算法的缺陷包括数据稀疏性、冷启动性及用户偏好各异等。其中，稀疏性是指用户对已有产品或服务方案评分很少，导致形成的数据矩阵稀疏。冷启动性是指在已有用户–项目规则库中，没有可以依据的新用户行为记录，难以对其推荐预测目标。协同过滤算法理论研究主要分为两种：基于模型的方法和基于内存的方法[1]。基于模型的 CFA 一般通过数据集训练出用户兴趣模型，并基于此模型对新用户进行推荐；而基于内存的方法则是通过建立用户–项目的评分矩阵，既可以预测用户缺失的数据，弥补数据稀疏性的问题，又能够向用户推荐相似的项目[2]。基于模型的方法构造模型的前期时间较长，但对用户的响应速度最快，而基于内存的方法可以节省模型构建时间，且注重内部数据的完整性。本文侧重于研究基于内存的 CFA，并融合了基于模型的 CFA 建模的思想，即结合两者的优势。

协同过滤算法具有稀疏性、冷启动及用户偏好不一致等问题，基于内存的 CFA 也有类似问题，学者们对其进行了改进或修正，主要包含以下两种情况：一是通过相似性度量的修改、最近邻的选择等提高推荐准确度；二是结合其他算法的 CFA 提高推荐效率，如结合矩阵分解、概率模型、贝叶斯模型及云模型等。针对第一种情况，文献[3]采用基于改进相似度的 CFA，构建网络项目的特征相似性模型，得到项目特征邻居并选择前 N 项推荐给用户，实现网络干扰的有效过滤，从而避免稀疏数据评价中产生的不足；文献[4]提出基于最近邻的 CFA，保证推荐精度提高的情况下，降低计算过程的复杂度；文献[5]根据用户历史行为数据差异性，提出基于改进用户相似度的算法，在相似度计算过程中添加平衡阈值，显著优化用户相似性并得到好的推荐结果。文献[4-6]主要采用基于最近邻的改进 CFA 方法，相互弥补最近邻和 CFA 两种方法的优劣势，进而提高推荐精度。如文献[6]在推荐算法中引入聚类方法，通过用户评分构建动态多维用户网络，采用改进 K-means 算法对用户聚类，将预测目标推荐给用户，形成应用于动态多维社会网络中的个性化推荐算法，从而提高推荐精度。上述文献中 CFA 的研究仅仅解决协同过滤中的数据稀疏性、冷启动、用户偏好不一致中的某项问题，并未全面考虑一次性解决这 3 个问题来提高推荐效率。本文综合考虑这 3 个问题，提出基于二部图的个性化推荐方法。首先，针对 CFA 的稀疏性问题，已有研究都是从弥补数据稀疏性的角度研究，文献[7]采用二部图算法，清理掉孤立的网页服务结点和工作流结点，提高数据完整性以及网页服务的推荐效率；文献[8-9]采用二部图与 CFA 结合的算法，解决数据稀疏性问题。考虑到二部图可以缓解 CFA 稀疏性，但合并数据精度不高，本文引入文献[10]中加权二部图的思想，合并关联度高的稀疏数据，形成完整性数据，提高数据处理的效率。其次，针对 CFA 的冷启动问题，相关文献研究采用基于内存的 CFA 预测目标，即改进自身算法，或是结合其他算法弥补 CFA 缺陷，而本文基于加权网络改进 CFA，对新用户预测推荐方案，提高预测目标的精确性，解决冷启动问题。最后，针对用户兴趣各异的问题，本文提出采用基于用户特征以及方案特征评分的协同过滤算法，不仅解决了用户偏好相异的问题，而且能够依据用户的简单需求提供合适的方案，即通过计算产品功能性特征的相似度，将相似度高的产品予以推荐。

本文采用加权二部图与加权网络的协同过滤混合算法，考虑用户偏好相异，将用户特征引入推荐系统，合理化资源分配过程，提高推荐精度。该方法的具体过程分为 3 个部分。首先，确

定用户的特征属性以及方案属性，对原始数据库中的用户特征采用加权二部图将用户进行分类，形成新的用户集；然后，采用加权二部图计算方案特征相似度，将相似度高的用户进行聚类，并采用 TOP-N 方法将相似用户进行排序，形成规则库；最后，采用协同过滤算法将新用户的特征与规则库中的用户特征进行匹配，并推荐已有的定制方案。

1 研究框架

假设企业中的数据库包含用户数据集和方案数据集，用户集$U=U_1,U_2,U_3,\cdots,U_m$，$U_i(1\leqslant i\leqslant m)$表示用户类型，用户定制产品的方案集$S=S_1S_2S_3,\cdots,S_n$，$S_j(1\leqslant j\leqslant n)$表示方案类型。用户集中的特征集为$F=f_1,f_2,f_3,\cdots,f_p$，其中，$f_g(1\leqslant g\leqslant p)$与 U 组成用户−特征的二维矩阵。方案集中的产品特征集为 $P=\{P_1,P_2,P_3,\cdots,P_t\}$，$P_a(1\leqslant a\leqslant t)$与 S_j 组成方案−特征的二维矩阵。本文是对用户与方案间的特征匹配，首先汇集用户已定制的方案，采用加权二部图，即确定用户−方案间的权重，根据用户相似特征分类，形成用户−方案规则库；然后采用基于加权网络的 CFA，对新用户$C=C_1,C_2,C_3,\cdots,C_k$推荐定制方案，$C_h(1\leqslant h\leqslant k)$表示单个用户，其所提需求特征$CP=CP_1,CP_2,CP_3,\cdots,CP_l$，$CP_q(1\leqslant q\leqslant l)$与用户−方案规则库中的用户特征进行相似度匹配，对方案集进行 TOP-N 排序并推荐，框架如图 1 所示。

2 基于加权二部图的用户−方案模型

为了提高个性化方案推荐方法提供可靠方案的准确率，本文提出采用加权二部图方法，合并高相似特征的用户，建立用户−多方案模型并对用户方案排序，建立用户−方案规则库，这是整个方案推荐过程中的核心部分。因此，采用 CFA 匹配新用户与规则库中的用户，推荐相应方案集，即如果有新用户需要定制方案，可以根据其描述特征，从已有用户−方案规则库提取用户特征匹配的规则，筛选适合的用户方案予以推荐。

2.1 二部图的概念

Guimera 提出的原始二部图，是指从随机图的期望值中，得出边缘集数的累积偏差。无向图 $G=(V,E)$，G 的每条边为 E，其顶点集合 V 可分为 2 个子集 U 和 S。

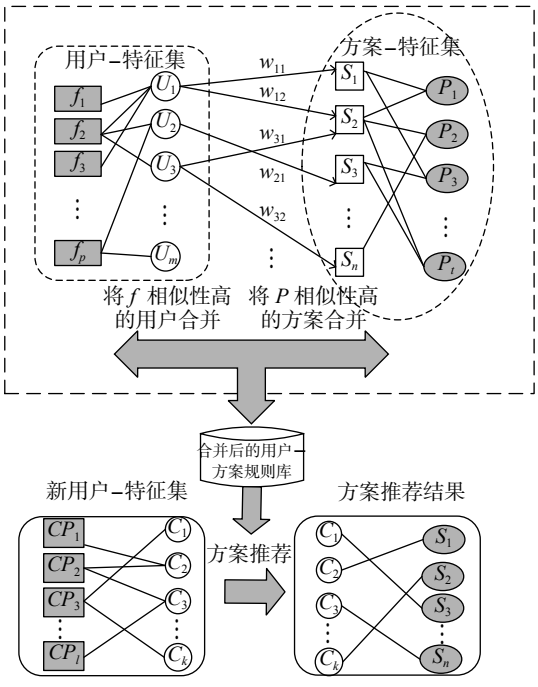

图 1 用户−方案评分二部图

Fig. 1 User-scheme score bipartite chart

定义 1[7] 二部图。

二部图是一类特殊的图，图中结点根据特性分为两类，其中，被归到同一类中的结点都具有相似的特性。边则连接来自不同类的 2 个结点，表示结点之间的关联。定义二部图

$$G=(U,S,E) \tag{1}$$

式中：U 为图的一类结点的集合，$U=\{U_i\}$，U_i 为一类结点集合的元素；S 为另一类结点的集合，$S=\{S_j\}$，S_j 为另一类结点集合的元素；E 为两类结点连接的边的集合，$E=\{e_{ij}\}$，e_{ij} 为集合的元素。

定义 2[7] 邻边。

二部图邻接矩阵 $\boldsymbol{A}$，$\boldsymbol{B}$ 用于表示二部图中的信息。a_{ij}，b_{ig} 为邻接矩阵中的元素。

$$a_{ij}=\begin{cases}1, & e_{ij}\in E\\ 0, & \text{其他}\end{cases}\qquad 0\leqslant i<m,\quad 0\leqslant j<n \tag{2}$$

$$b_{ig}=\begin{cases}1, & e_{ig}\in E\\ 0, & \text{其他}\end{cases}\qquad 0\leqslant i<m,\quad 0\leqslant g<p \tag{3}$$

a_{ij} 数值用于判断结点间是否有连接，即用户 U_i 有无选择方案 S_j，若选择，则 a_{ij} 值为 1，反之值为 0，m 和 n 分别表示二部图中两类结点的数量。b_{ig} 是指用户 U_i 有无描述特征 f_g，若有，则该值为 1，反之为 0。

定义 3[7] 权重。

边的权重是指边两端的结点间关联的重要程度，用户−方案之间的权重为w_{ij}，表示方案 S_j 在 U_i 所有方案中评价值的比重，即用户对方案的偏好度，w_{ij}越大，表示用户对方案越认可，则用户和方案之间的关联性越强。

$$w_{ij}=\frac{r_{ij}}{\sum_{i=1}^{n} r_{ij}} \tag{4}$$

式中，r_{ij} 表示用户 U_i 对方案 S_j 的评分。

定义 4　评分矩阵。

方案−特征间的评分表示为 SP_{nt}，是指特征在方案中的重要程度。用户 U_i 对非功能性特征 f_g 评分结果为$d_{U_{ip}}$，用户对每个特征的认可程度从差到好表示为 1−5，0 表示用户未选择该特征，评分结果形成二维矩阵 $\boldsymbol{B}$。

$$\boldsymbol{B}=\begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_p \end{array}\begin{array}{c} \begin{array}{cccccc} U_1 & U_2 & U_3 & \cdots & U_n \end{array} \\ \begin{bmatrix} 1 & 4 & 2 & \cdots & 4 \\ 0 & 0 & 2 & \cdots & 0 \\ 5 & 2 & 0 & \cdots & 3 \\ \vdots & \vdots & \vdots & & \vdots \\ 4 & 4 & 2 & \cdots & 3 \end{bmatrix} \end{array} \tag{5}$$

如果用户没有选择任何特征标签，可以根据其定性描述，将其按专家打分重要性进行分类，转化成定量化数据，从而解决模糊不确定的评价信息。

定义 5　评价指标。

针对定性化描述的评价指标，具有模糊不确定性，而式(6)通过定量化方式来计算方案分配给用户的比重。本文通过专家评分将定性指标转为定量数据，采用式(6)计算出具体评分指标值。

$$R=\frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{6}$$

式中：R 表示适合用户 U_i 的方案 S_j 排名计算方式；w_{ij}表示方案 S_j 在 U_i 所有方案中评分值的比重，$\sum_{j=1}^{n} w_{ij}$为方案 S_j 被采用后被用户评价的分值总和。

定义 6[10]　二部图相似度。

当新用户选择方案时，根据其所描述的功能性或非功能性特征，计算新用户 U_s 特征 f_g 与规则库中用户 U_i 特征相似度为 sim_{CU}。

$$sim_{CU}=\sum_{g=1}^{p}\frac{b_{ig}b_{sg}}{d_{\alpha}(f_g)}/\min\{d(U_i),d(U_s)\} \tag{7}$$

式中：b_{ig}，b_{sg} 表示用户 U_i 和 U_s 是否选择特征 f_g，$d(u_i)$，$d(U_s)$ 表示用户 U_i 和 U_s 对特征 f_g 的评价值；$d_{\alpha}(f_g)$是用户 U_i，U_s 对共同评分特征的加权值。

2.2　基于加权二部图的用户−特征聚类过程

考虑到新用户先前没有方案推荐行为记录,难以计算该用户和其他用户的相似度,即推荐算法的冷启动问题,也是目前推荐系统面临的关键问题。所以，在用户−方案规则基础上引入特征标签，形成用户−特征的推荐方案模型。采用加权二部图对用户特征分类，用户将已有的特征评分矩阵投影在二部图中，通过式(6)对用户非功能性特征进行聚类，形成新用户集$U'=\{U'_1,U'_2,U'_3,\cdots,U'_b\}$，这样既减少冗余数据，节省存储空间，又便于个性化方案的查询和推荐，具体如图 2～4 所示。

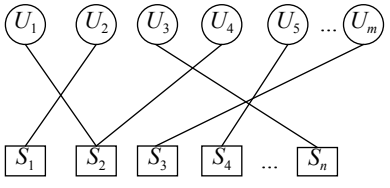

图 2　用户−方案初始集

Fig.2　User-scheme initial set

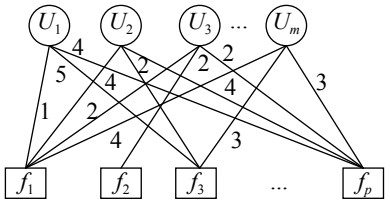

图 3　用户−特征评分

Fig.3　User-feature score

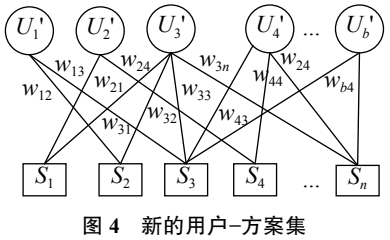

图 4　新的用户−方案集

Fig.4　New user-scheme set

首先取出用户对特征的评分，其次利用式(6)求出用户特征的相似度，将相似性高的近邻用户对应方案以及用户−方案间的权重合并，形成一对多的方案集，既能压缩用户集，节约存储空间，又能根据用户−方案规则库中的用户特征匹配方案推荐，给予新用户更多选择推荐方案的机会。

2.3　基于加权二部图的方案过滤推荐

依据形成的新用户−方案集，针对局部用户对

应的方案集太多，如果都推荐给用户，难以保证满意度且效率低。对此，采用二部图的邻居元素、权重以及二部图评价指标的计算方式，将同一用户对应方案集由高到低排序，选取前 N 项的方案推荐，具体的过程用伪代码描述如下：

Input User $U=\{U_i\}, V=\{V_j\}$

count $sim(U_i,V_j)$

If $sim(U_i,V_j)>0.8$

combined U_i and V_j

formed new sets $U'=\{U'_1,U'_2,U'_3,\cdots,U'_b\}$

Then count w_{ij} in S_n *of* U' and sort

count $sim^N(U',C)$ between f_g of U' and New user C

Recommend w_{ij} in S *of* U'

现介绍具体步骤。

步骤 1 首先通过以往用户方案的评价信息，计算w_{ij}并输入到二部图中，连接 2 个二部图构建用户−方案集，方案−特征集如图 5 所示。SP 表示方案特征的评分。

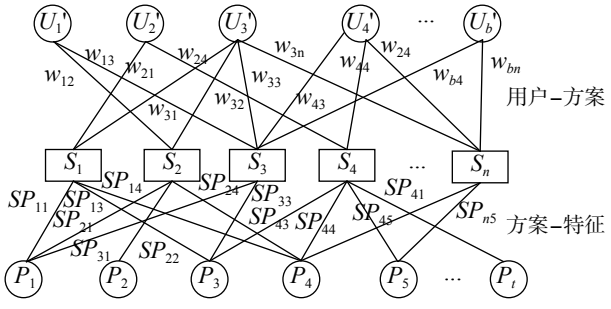

图 5 二部图组合

Fig.5 Combination of bipartite graphs

步骤 2 遍历二部图的结点，如用户结点 U_1'对应的方案结点为 S_2 和 S_3，这是用户−方案集的二部图，方案 S_2 对应特征为 P_1，P_2，P_4，方案 S_3 对应特征为 P_1，P_3，这是方案−特征的二部图。本步骤计算方案−特征之间的关系，先计算特征 P 在方案中对应的比重，如 P_1 在 S_2 的比重为 $\frac{SP_{21}}{SP_{11}+SP_{21}+SP_{31}}$，$P_2$，$P_4$ 同理，得出加权分

$$S_2\_P=\frac{SP_{21}}{SP_{11}+SP_{21}+SP_{31}}a_{21}+SP_{22}a_{22}+\frac{SP_{24}}{SP_{14}+SP_{24}+SP_{44}+SP_{n4}}a_{24}$$

同理，

$$S_3\_P=\frac{SP_{31}}{SP_{11}+SP_{21}+SP_{31}}a_{31}+\frac{SP_{33}}{SP_{13}+SP_{33}+SP_{43}}a_{33}$$

式中：a_{21} 表示 S_2 与 P_1 之间的评分；a_{22}，a_{24}，a_{31}，a_{33} 同理。

步骤 3 计算方案 S_2 和 S_3 在用户结点U_1'的比重，对$U_1'_S_2=w_{12}\times S_2_P$，$U_1'_S_3=w_{13}\times S_3_P$的值排序，筛选自己所需的方案。

通过以上步骤，本文对特征相似的用户聚类及对应方案排序和推荐，形成用户−方案规则库，便于给新用户推荐已有定制方案，提高方案推荐的效率。

3 改进的协同过滤算法

基于上述已形成的用户−方案规则库，采用加权网络的协同过滤算法，计算新用户特征与用户−方案规则库中用户特征的相似度，给新用户推荐方案。如果与规则库中的某用户特征相似度高，则将该用户对应方案给予推荐。即使规则库中没有新用户定制方案的行为记录，也能推荐方案，从而解决 CFA 的冷启动问题。协同过滤算法是个性化推荐技术中最为成熟的算法，针对 CFA 稀疏性和冷启动问题，文献[11]采用基于加权网络的推荐算法予以处理，加权网络的连接强度通过拓扑结构不仅弱化了稀疏性问题，而且根据用户评分帮助其推荐未曾产生行为的方案，解决冷启动问题，因此，引入基于加权网络的相似度连接方法。传统 CFA 主要通过收集新用户偏好，匹配已有相似用户集，将用户集对应方案予以推荐，因此，该算法核心在于相似度计算。目前广泛采用的相似系数主要包括皮尔森系数(Pearson correlation coefficient, PCC)、Jaccard 系数、余弦系数等。皮尔森系数是基于协方差标准化和无量纲化，既能判断 2 个随机变量的线性关系，又能得到相关性指标；Jaccard 系数侧重于计算 2 个集合间的相似度，且能处理非对称的二元变量，因而得到的结果并非同等重要；余弦系数一般用来计算向量间的相似度，侧重空间多维方向性，如当下网页的爬虫技术，通过计算网页中相关内容向量的相似度，删除重复网页。本文 CFA 主要侧重于用户−方案特征以及用户集评分的相似度计算，采用皮尔森系数和 Jaccard 系数，分别用于计算用户−方案间的评分相似度和项目间的评分相似度。

3.1 相关概念

皮尔森系数常用于推荐系统中的相似性计算，确定非功能属性值的相似性。式(8)为皮尔森相关系数的计算方法。

$$sim(U,V)^{\mathrm{PCC}}=\frac{\sum_{f\in I_u\cap I_v}(r_{uf}-\overline{r_u})(r_{vf}-\overline{r_v})}{\sqrt{\sum_{f\in I_u\cap I_v}(r_{uf}-\overline{r_u})^2}\sqrt{\sum_{f\in I_u\cap I_v}(r_{vf}-\overline{r_v})^2}} \quad (8)$$

式中：I_u 和 I_v 分别表示用户 U 和 V 评价过的非功能性特征 F 的集合；r_{uf}和r_{vf}分别表示用户 U 和 V 对非功能性特征的评分；$\overline{r_u}$和$\overline{r_v}$分别表示用户 U 和 V 对共同评价非功能性特征 F 的平均值。所以，用户相似度的数值属于闭区间[−1, 1]，值越大表示用户 U 和 V 越相似。

式(9)为 Jaccard 系数的计算方法。

$$J(U,V)=\frac{\sum r_u\cap r_v}{\sum r_u\cup r_v} \quad (9)$$

式中，r_u，r_v分别表示用户 U，V 对所有非功能性特征 F 的评分集。

3.2　基于加权网络的协同过滤算法

为了解决新用户的方案定制推荐问题，建立一个加权网络的评分项目，其中，每条边的权重值为共同评分用户的数量。此外，为了预测推荐目标的精确性,本文通过阈值限制匹配用户，所以，在加权网络中，2 个用户必须满足其共同评分项的数量大于阈值β这个条件，从而过滤掉权重低于阈值β的用户集[11]。

$$I_{U-V}(\beta)=\sum_{o\in f}\frac{si(w_{uo}\cap w_{vo})}{si(w_{uo}\cup w_{vo})} \quad (10)$$

式中：I_{U-V}表示用户 U 和 V 的连接强度；β为对特征共同评分项的数量设置的阈值；w_{uo}表示用户 U 对 O 特征评分的方案编号集；w_{vo}表示用户 V 对 O 特征评分的特征编号集；si 表示特征数目集。

由于皮尔森相关系数主要在求解 2 个用户之间评分相似度起作用，Jaccard 系数侧重于解决项目的评分相似度方面的问题，但是，都存在数据稀疏性问题，导致目标预测结果不精确。而连接强度通过拓扑结构弱化了稀疏性问题，一定程度上提高了预测准确率。综合考虑以上 3 种方法，本文提出采用新用户相似度计算公式：

$$sim_f(U,V)=sim(U,V)^{\mathrm{PCC}}\cdot J(U,V) \quad (11)$$

$$sim^N(U,V)=sim_f(U,V)\cdot I_{U-V}(\beta) \quad (12)$$

式中：$sim_f(U,V)$是由传统的相似度组成，表示用户 U 和用户 V 的相似度；$sim^N(U,V)$ 表示经过阈值筛选后的综合相似度；$I_{U-V}(\beta)$为连接强度，用来改善传统相似性的不足，弥补了稀疏性问题的不足。

采用上述新相似度计算方法对新用户 $C=\{C_1,C_2,C_3\cdots,C_k\}$,$C_h(1\leqslant h\leqslant k)$所提的某些需求特征 $CP=\{CP_1,CP_2,CP_3,\cdots,CP_l\}$(即 $CP_q(1\leqslant q\leqslant l)$) 与方案规则库中的用户$U'=\{U'_1,U'_2,U'_3,\cdots,U'_b\}$中的特征 F 匹配，若计算的$I_{U'-C}(\beta)$低于给定阈值β，或者 Jaccard 系数低于给定阈值，则越过当前用户，与下一个用户匹配。反之，计算相似度$sim^N(U',C)=sim_f(U',C)\cdot I_{U'-C}(\beta)$，并采用 TOP-N 方法对相似度排序，取相似度最大的用户所对应方案进行推荐，这样能使用户在不了解产品功能的情况下，根据其所提的非功能性需求，给予合理的推荐方案，从而解决 CFA 冷启动的问题，提高推荐效率。

4　实例研究及对比分析

4.1　实例研究

某汽车 4S 店想要提高销售的效率，希望定制产品的个性化推荐方案，能够准确地推送给新用户，满足其需求。与文献[12]建立用户偏好模型不同，本文建立用户–特征和方案–特征 2 个模型，并用二部图建立 2 个模型间的联系。通过整理以往顾客购买产品历史记录信息，本文主要选取用户购买实用性的家庭装车型，选择代表性强的数据，得出用户特征包括: 变速箱 f_1，座位数 f_2，经济环保 f_3，安全性 f_4，舒适性 f_5，售后服务 f_6，成本 f_7，用户 $U=\{U_1,U_2,U_3,\cdots,U_m\}$,$U_i(1\leqslant i\leqslant m)$选择的方案包括：小型车 S_1，S_2，S_3，SUV 车型 S_4，S_5，S_6，MPV 车型 S_7，S_8。方案特征包括：动力性 P_1，燃油经济性 P_2，制动性能 P_3，汽车通过性 P_4，轮胎性能 P_5，载客数 P_6，汽车噪音 P_7，排放性能 P_8，安全性 P_9。具体的指标如表 1 和表 2 所示。

该企业的汽车销售数据库包括用户方案对应的非功能性和功能性特征的数据。首先将非功能性数据，即用户特征集取出，设置相似度阈值为 0.9，并基于用户特征计算相似度，超过给定阈值则合并用户对应的方案集；其次采用基于二部图排序的方法，筛选并得到某用户的多方案，筛选前 N 项的方案给予推荐；最后采用改进 CFA，对新用户 $C=\{C_1,C_2,C_3,\cdots,C_k\}$的某些需求特征 $CP=\{CP_1,CP_2,CP_3,\cdots CP_l\}$(即$CP_q(1\leqslant q\leqslant l)$) 与生成的方案规则库匹配，现介绍具体的步骤。

步骤 1　根据汽车购买用户特征，f_2，f_3，f_4，f_7 由低到高用 1–5 表示，f_1，f_5，f_6 由低到高用 1–4 表示。采用式(6)计算特征数据相似度，但如果用户并未对特征打分，依据其定性描述，采用专家已评定好的标准将其转化为定量化数据。采用二

表 1 用户特征指标

Tab.1 User characteristics indicators

变速器 f_1	座位数 f_2	经济环保 f_3	安全性 f_4	舒适性 f_5	售后服务 f_6	成本 f_7
手动	7	耗油少	高	心理舒适性	维修	<10 万元
自动	6	耗油较少	较高	便利性	保养	10～20 万元
手自一体	5	耗油一般	一般	人机交互	定期检查	20～30 万元
双离合	4	耗油较多	较差	智能化	折旧	30～50 万元
	2	耗油多	差			>50 万元

表 2 方案特征指标

Tab.2 Programme characteristic indicators

动力性 P_1	燃油经济性 P_2	制动性能 P_3	汽车通过性 P_4	轮胎性能 P_5	载客数 P_6	汽车噪音 P_7	排放性能 P_8	安全性 P_9
加速时间	市区	制动效能	离去角	差	2	低	<1.3 L	膝部气囊
最高速度	郊区	抗衰退性能	最小转弯半径	中	4～5	中	1.3～1.6 L	头颈保护系统
爬坡能力高	综合耗油	方向稳定性	越障能力	防抱死好	6	较高	1.7～3.0 L	安全气囊

部图相似度式(6)，计算用户评价汽车特征的数据，如图 6 所示，其中，权重α设置为 0.5。

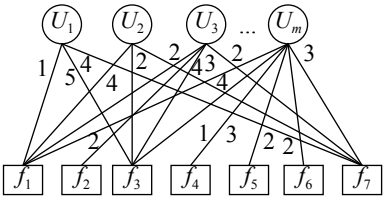

图 6 用户–特征图

Fig.6 User-feature chart

以 U_1 和 U_3 的相似度计算为例，$sim_{U_1U_3}=\frac{1\times1}{\sqrt{1}\sqrt{2}}/1+\frac{1\times1}{\sqrt{5}\sqrt{3}}/4+\frac{1\times1}{\sqrt{4}\sqrt{2}}/2\approx0.946\,8$，超过阈值 0.9，因此，合并 U_1 和 U_3 的用户特征，将 U_1 和 U_3 定制的方案聚集并组成U_1'。同理，合并用户之间特征相似度超过 0.9 的方案，并以用户共同评价特征的最大值作为新用户评分值，相异的评价特征不变并保留，得出用户–方案集 $U'=\{U_1',U_2',U_3',\cdots,U_b'\}$，其中，用户方案权重$w_{ij}$由用户对方案评分的比值决定，并代入图 7，具体过程如图 8 所示。

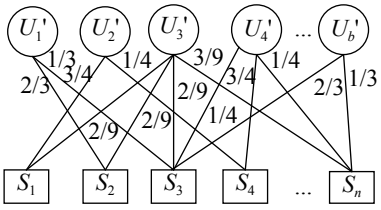

图 7 合并后的用户–方案集

Fig. 7 Merged users-scheme set

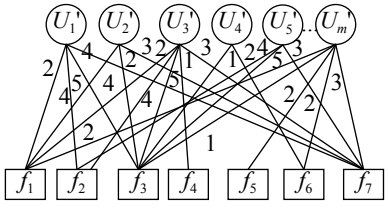

图 8 合并后的用户–特征集

Fig. 8 Merged user-feature set

步骤 2 采用加权二部图对每个购买汽车的用户其对应的方案集排序，即根据方案特征的评分，通过 1，2，3 来表示方案特征的好坏，采用加权二部图的方法计算并排序，选取排名靠前 N 项的方案推荐，具体如图 9 所示。

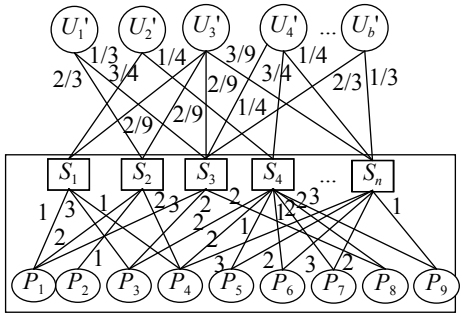

图 9 用户方案–特征集

Fig.9 User scheme-feature set

对图 9 框线部分的方案–特征排序，如对用户 U_3'下的方案 S_1，S_2，S_3，S_n 的排序，首先根据式(5)计算 S_1 的评分值，其对应的功能特征 P_1，P_3，P_4 的比重分别为 1/(1+2+3)=1/6，3/(3+2+2)=2/7，

1/(1+2+2+1)=1/6，得出 S_1_P=1/6+2/7+1/6=13/21，同理，得出 S_2_P=5/3，S_3_P=9/7，S_n_P=19/10。而用户对方案评分值，是通过权重和方案比重值的乘积公式进行计算，结果得出 $U'_3\_S_1 = 2/9\times 13/21 \approx 0.138$，$U'_3\_S_2 = 2/9\times 5/3 \approx 0.37$，$U'_3\_S_3 = 2/9\times 9/7 \approx 0.286$，$U'_3\_S_n = 3/9\times 19/10 \approx 0.633$。由此得出结论，对于用户 U'_3来说，依据加权求和的评分值大小对方案排序，由高到低分别为 S_n，S_2，S_3，S_1。因此，方案 S_n 是推荐给用户的首选，即越障能力高、轮胎防抱死性能好、汽车噪音中等、能保护头颈，如 SUV 类型的车推荐给该用户。与此类似，当其他用户对应特别多的方案集时，采用此方法排序筛选，可以去除低价值的方案，提高推荐效率，从而生成新方案规则库。

步骤 3　基于步骤 2 生成的方案规则库，给欲购买汽车的新用户提供便利。当新用户进入 4S 店购买汽车时，让其提出需求汽车的非功能性特征 F，采用 1−5 这 5 个等级评分。据此，本文采用 CFA 方法对相似度进行改进，将 Jaccard 系数阈值设置为 0.8 以及连接强度阈值 β都为 0.6，如果计算出的连接强度 $I_{U-V}(\beta)$或者 Jaccard 系数低于这个阈值，直接省去从用户−方案规则库中匹配的过程；如果满足这个阈值范围，则计算新相似度并用 TOP-N 算法排序，找出相似性最大的用户所对应的方案集并推荐，如图 10 所示。

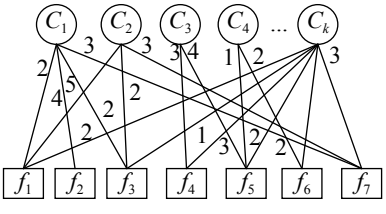

图 10　新用户−特征集

Fig.10　New user-feature set

以新用户 C_1 特征匹配过程为例，将新用户 C_1 与规则库中的用户 U'_1的非功能性特征匹配，采用式（7）～（11）计算，得出 J=11/18，$I_{U'C-f}$ =4/5。$I_{U'C-f}$虽然满足给定的阈值 0.6，但是，J不满足给定值 0.8，所以，省略此匹配用户，与其他用户特征重新匹配。将 C_1 与U'_3匹配，I_{U_1C-f} =4/5，J=13/14，都符合给定的阈值，则继续计算 $sim_{U_3'-C_1}{}^{\mathrm{PCC}}$=1，所以，$sim_{U_3'-C_1}{}^{N}$=26/35。同理，根据特征计算其他用户与新用户 C_1 的相似度，如$I_{U_5'C-f}$=3/4，J=12/14，$sim_{U_5'-C_1}{}^{N}$=9/14 等，并排序。由此得出 U'_3与新用户 C_1 的相似性最大，从而将U'_3经过筛选后的方案推荐给 C_1，U'_3的非功能性特征，包含汽车变速箱自动、耗油多、安全高、价位在 20～30 万左右。据此，其所对应的方案 S_n，汽车功能必须是越障能力高、轮胎防抱死性能好、汽车噪音中等、能保护头颈，如 SUV 类型的车可以推荐给用户。由此可见，本文的算法可以满足潜在的或者不熟悉产品功能特征的但又想获取性价比高的方案的用户需求，既能帮助企业提高收益，又能提高方案推荐效率。

4.2　算法性能的对比分析

针对协同过滤算法的缺点，考虑数据稀疏性、冷启动及用户偏好不一致这 3 个问题，采取基于加权二部图的个性化推荐方法来解决。加权二部图方法是通过计算用户−方案相似度，形成用户−方案规则库。推荐方案分为采用和未采用两种形式，采用方案分为规则库推荐和未推荐，分别用 tp 和 fp 表示；未采用方案分为规则库推荐和未推荐，分别用 tn 和 fn 表示。为了证明该算法的优越性，本文提出采用 F-Score 评价体系[13]，衡量该算法与传统协同过滤预测目标的性能，主要包含准确率 P，召回率 R 以及 F-Score 评价体系F'，具体定义为

$$P = \frac{tp}{tp+tn} \tag{13}$$

$$R = \frac{tp}{tp+fn} \tag{14}$$

$$F' = \frac{PR}{P+R} \tag{15}$$

准确率和召回率是用于衡量该算法构建规则库的覆盖率以及推荐精度，而 F-Score 是衡量这两者间的综合指标。本文以对某汽车 4S 店的方案的问卷调查数据为例，根据传统协同过滤算法、已有的加权二部图算法以及本文算法对 Top-N 中的前 100，200，300，400，500 项方案，分别计算评估体系的参数的总体平均值，结果如图 11 所示。

由图 11 可知，本文算法相比较协同过滤算法、加权二部图算法而言，优越性很大。如在前 300 项的数据下，传统 CFA 的综合 F-Score 评价指标为 0.47，加权二部图的 F-Score 评价指标为 0.60，而本文算法 F-Score 评价指标为 0.783，明显高于其他 2 个算法。在方案数较少的情况下，CFA 和加权二部图的 F-Score 评价指标值不相上下，直到方案数在 200 之后，综合指标的差距才开始拉大，并且加权二部图推荐性能明显优于 CFA。从图 11 中可知，随着方案数的增加，相比

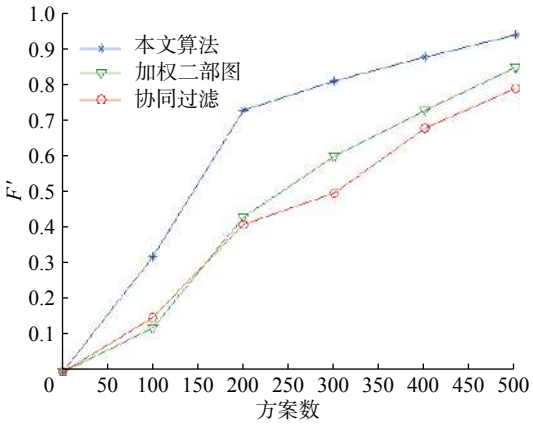

图 11 不同算法 F-Score 比较结果

Fig. 11 F-Score comparison results of different algorithms

较 CFA、加权二部图的推荐效果，本文算法的推荐精度性能优势越明显，说明在推荐目标方案方面，本文算法更加有效，进一步证明本文算法能够更好地预测新用户的目标方案。

5 结束语

针对协同过滤算法执行过程中存在数据稀疏性、用户兴趣各异以及冷启动问题，提出基于加权二部图的个性化推荐算法予以解决，提高了推荐效率。本文算法的优点如下：

a. 采用二部图方法计算用户特征的相似度，将相似度高的用户聚类，形成一对多的用户方案集，节约了数据库存储空间，减轻了数据稀疏性。

b. 针对聚类后的用户方案集，根据用户对方案特征的评分，采用加权二部图计算方案评分值，并用 TOP-N 排序，保留前 N 项方案，解决了用户兴趣各异的情况，为新用户推荐不同的方案。

c. 对于零购买记录的新用户，本文采用改进相似度，即将皮尔森系数、Jaccard 系数以及连接强度相结合，如对 Jaccard 系数以及连接强度设定阈值，相似度不满足该阈值，直接省略推荐过程，反之，将与新用户相似性高的用户方案予以推荐。该方法解决了以往推荐算法的冷启动问题，提高了算法执行时间以及推荐效率。

以上是本文算法的优点，但对于小型数据库而言，方案相似度 TOP-N 排序方法效果不很明显。此外，本文算法在实例分析中，也只是展示方案了解和实施过程，即将用户–方案库中的方案推荐给新用户，并未针对用户某些特殊需求，定制特殊方案，这也是本文下一步想拓展的研究方向。

参考文献：

[1] SHI Y, LARSON M, HANJALIC A. Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges[J]. ACM Computing Surveys, 2014, 47(1): 3.

[2] ZHANG J, LIN Y J, LIN M L, et al. An effective collaborative filtering algorithm based on user preference clustering[J]. Applied Intelligence, 2016, 45(2): 230–240.

[3] 贾忠涛, 吴颖川, 刘志勤. 一种协同过滤算法在网络干扰过滤中的应用[J]. 计算机仿真, 2016, 33(1): 284–287.

[4] 罗辛, 欧阳元新, 熊璋, 等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报, 2010, 33(8): 1437–1445.

[5] CHEN H, LI Z K, HU W. An improved collaborative recommendation algorithm based on optimized user similarity[J]. The Journal of Supercomputing, 2016, 72(7): 2565–2578.

[6] 王鑫, 黄忠义. 网络资源中基于 K-Means 聚类的个性化推荐[J]. 北京邮电大学学报, 2014, 37(S1): 120–124.

[7] 姜波, 张晓筱, 潘伟丰. 基于二部图的服务推荐算法研究[J]. 华中科技大学学报(自然科学版), 2013, 41(S2): 93–99.

[8] 李霞, 李守伟. 面向个性化推荐系统的二分网络协同过滤算法研究[J]. 计算机应用研究, 2013, 30(7): 1946–1949.

[9] HU X H, MAI Z C, ZHANG H L, et al. A hybrid recommendation model based on weighted bipartite graph and collaborative filtering[C]//Proceedings of 2016 IEEE/WIC/ACM International Conference on WEB Intelligence Workshops. Omaha, NE, USA: IEEE, 2016: 119–122.

[10] CHEN D E, YING Y L. A collaborative filtering recommendation algorithm based on bipartite graph[J]. Advanced Materials Research, 2013, 756-759: 3865–3868.

[11] ZHAN W P, LI Q. An improved collaborative filtering based on a weighted network and triadic closure[C]// Proceedings of 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops. Beijing, China: IEEE, 2015: 1760–1763.

[12] GAN M X, JIANG R. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities[J]. Decision Support Systems, 2013, 55(3): 811–821.

[13] ZHANG J, PENG Q K, SUN S Q, et al. Collaborative filtering recommendation algorithm based on user preference derived from item domain features[J]. Physica A: Statistical Mechanics and Its Applications, 2014, 396: 66–76.

（编辑：石 瑛）