·人工智能及识别技术· 文章编号:1000-3428(2017)10-0160-07 文献标志码:A 中图分类号:TP391

基于动态标签偏好信任概率矩阵分解模型的推荐算法

杨亚东,熊庆国

(武汉科技大学 信息科学与工程学院,武汉 430081)

摘 要: 为提高推荐算法性能,解决数据稀疏和冷启动因素造成的推荐精度不高的问题,提出一种改进的协同过滤推荐算法。基于三元组表示形式,利用标签集、用户集和项目资源集构建标签、用户以及项目之间的动态联系,并进行信任值评分矩阵的计算,使用信任评分矩阵融合协同推荐过程,构建概率矩阵分解模型,并基于期望最大法进行模型的求解。实验结果表明,与采用基于余弦、皮尔逊相关系数和启发式相似度模型的算法相比,该算法具有较低的绝对误差均值以及较高的覆盖率、精度与召回率。

关键词: 协同过滤推荐;数据稀疏;冷启动;概率矩阵分解;标签偏好;期望最大法

中文引用格式: 杨亚东,熊庆国. 基于动态标签偏好信任概率矩阵分解模型的推荐算法[J]. 计算机工程,2017,43(10):160-166.

英文引用格式: YANG Yadong,XIONG Qingguo. Recommendation Algorithm Based on Dynamic Label Preference Trust Probability Matrix Decomposition Model[J]. Computer Engineering,2017,43(10):160-166.

Recommendation Algorithm Based on Dynamic Label Preference Trust Probability Matrix Decomposition Model

YANG Yadong,XIONG Qingguo

(School of Information Science and Engineering,Wuhan University of Science and Technology,Wuhan 430081,China)

【Abstract】 In order to improve the collaborative performance of recommendation algorithms and solve the problem of low recommendation accuracy caused by sparse data and cold start,an improved collaborative filtering recommendation algorithm is proposed in this paper. Based on three tuple representation,it uses the tag set,the user set and the project resource set to construct the dynamic relationship among the labels,the users and the project,and it also computes the trust value score matrix. It uses the trust rating matrix fusion collaborative recommendation process to construct probability matrix decomposition model and solves the model the expectation maximization method. Experimental results show that compared with other algorithms which are based on cosine,Pearson correlation coefficient and heuristic similarity model,this algorithm has lower absolute mean error as well as higher coverage rate,precision and recall rate.

【Key words】 collaborative filtering recommendation; data sparse; cold start; probability matrix decomposition; label preference; expectation maximization method

DOI: 10.3969/j.issn.1000-3428.2017.10.027

0 概述

随着互联网的快速进步,可提供的信息资源也日渐丰富,传统算法无法快速为用户提供所需的感兴趣信息,信息的获取效率不高[1-2]。推荐系统与传统形式的搜索引擎存在差异,其目的是为用户进行信息的过滤[3]。推荐系统最常用算法是协同过滤推荐,该算法可对用户偏好进行过滤推荐。

目前,有研究称亚马逊 P2P 网站的商品过滤推荐使得网站的营收增长近 35%。但是协同推荐算法也存在通用的问题,即数据稀疏和冷启动会造成推荐精度不高[4]。近年来,社会化属性的网络得到广泛研究,并在推荐系统算法设计过程中引入了信任关系,提出了基于信任关系的过滤推荐系统[5]。例如,文献[6]给出基于信任关系的协同过滤推荐系统设计原理,并指出与传统意义上采用的基于相似度评价的协同过滤算法模型相比,信任关系模型在进行参考用户的选取上存在差异。信任关系模型使得用户能够充分利用其信任用户的信息,并通过信息处理实现信息的高效获取[7]。利用用户之间存在的社会网络可提高信息推荐的覆盖率和质量,

基金项目: 湖北省教育厅科研计划项目(Q20151101);赛尔网络下一代互联网技术创新项目(NGII20150301)。

作者简介: 杨亚东(1990—),男,硕士研究生,主研方向为人工智能;熊庆国,教授。

收稿日期: 2016-09-13 **修回日期:** 2016-10-26 **E-mail:** yangydongls@qq.com

实现数据稀疏问题的有效缓解,近段时间以来引起了学者的广泛关注。

推荐系统的信任模型,同时也存在一些不足[8-9],如覆盖率指标与推荐精度的互斥问题,由于采用的社会网络信任取值为二进制形式,即 0-1 表示方式,会导致推荐精度下降;二进制 0-1 表示方式,存在用户间信息的不对称性,因此,信任值的合理计算对于提高算法性能至关重要。

以上述研究为基础,本文提出一种利用用户标签偏好间的动态联系计算的信任评分矩阵方法,构建信任概率矩阵分解模型(Trust Probability Matrix Decomposition Model,TPMDM),并在此基础之上对协同过滤推荐算法进行改进。

1　概率矩阵分解的动态标签信任模型

综合考虑潜在用户特征与信任之间存在的内部关联,并通过概率矩阵分解实现信任评分矩阵的预测和计算。

1.1　信任概率矩阵分解模型

信任概率矩阵分解模型主要源自对 SocialMF 和概率矩阵分解模型(Probability Matrix Decomposition Model,PMDM)模型的综合与改进[10-11]。给出上述文献所使用概率矩阵分解模型,如图 1 所示。

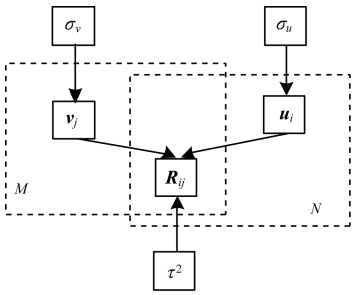

图 1　概率矩阵分解模型

图 1 所示为文献[10-11]中所采用的概率矩阵分解模型,该模型为具有线性概率特征的矩阵分解模型,将用户对项目评价问题变为概率问题。但是该模型存在的问题是其未考虑对参与评分用户的信任级别。实际上,不同用户在评分过程中具有不同的角色和定位,过于均值化的评分概率预测,不利于充分利用重点用户的评分意见,获得更为合理的决策结果,因此,为解决该问题,本文融合信任模型,提出如下 TPMDM 信任评价模型,如图 2 所示。

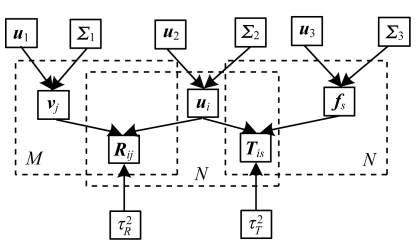

图 2　TPMDM 模型示意图

TPMDM 模型是在评分矩阵的分解过程中,充分考虑用户信任评分情况,对信任矩阵和评分矩阵进行高斯先验分布的多元化处理。在图 2 所示模型中,假设评分矩阵为 N 行 M 列,每行 i 对应特征为 $\boldsymbol{u}_i$,而每列 j 对应特征为 $\boldsymbol{v}_j$,$\boldsymbol{u}_i$ 和 $\boldsymbol{v}_j$ 特征元素服从高斯特征分布,形式为 $N(\boldsymbol{v}_i|0,\sigma_v^2)$ 与 $N(\boldsymbol{u}_j|0,\sigma_u^2)$,且项目概率分布满足独立特性。图中,矩阵 $\boldsymbol{R}_{ij}$ 可利用高斯特征分布 $N(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j,\tau_R^2)$ 获得,这样项目的用户评分可利用概率问题进行表示。基于 $\{\boldsymbol{u}_1,\boldsymbol{\Sigma}_1,\boldsymbol{u}_2,\boldsymbol{\Sigma}_2,\boldsymbol{u}_3,\boldsymbol{\Sigma}_3\}$ 的高斯协方差对角分布可实现潜在特征 $\boldsymbol{u}_i$ 与 $\boldsymbol{v}_j$ 的回归,该过程的损失函数为:

$$\begin{aligned}L(\boldsymbol{R},\boldsymbol{T},\boldsymbol{U},\boldsymbol{V}) = &\frac{1}{2}\sum_{u=1}^{N}\sum_{i=1}^{M}I_{ui}^{R}(\boldsymbol{R}_{ui}-g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i))^2\\&+\frac{\lambda_T}{2}\sum_{u=1}^{N}((\boldsymbol{U}_u-\sum_{V\in Nu}\boldsymbol{T}_{u,v}\boldsymbol{U}_v)^{\mathrm{T}}\\&\cdot(\boldsymbol{U}_u-\sum_{V\in Nu}\boldsymbol{T}_{u,v}\boldsymbol{U}_v))\\&+\frac{\lambda_U}{2}\sum_{u=1}^{N}\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{U}_u+\frac{\lambda_v}{2}\sum_{i=1}^{M}\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{U}_i\end{aligned}\qquad(1)$$

其中,$\boldsymbol{R}$ 代表评分矩阵;$\boldsymbol{T}$ 代表信任矩阵;$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 代表特征矩阵;$g(\cdot)$ 是正则化逻辑映射;λ_v,λ_u 与 λ_T 代表正则化参数(分别对应项目、用户与信任),可利用梯度下降对潜在特征逻辑模型进行优化。

1.2　动态标签关系

标签设置的主要是用户在项目上添加的注解,带有一定的社会属性。对于用户浏览、组织和推荐资源,具有一定的引导和帮助作用。首先,构建通用概念模型,选取三元组表示方式[12],如图 3 所示。利用标签集(tags)、用户集(users)和资源项目集(items),构建标签(T_i)、用户以及资源项目之间存在的动态联系。用户集表征的是用户空间,其涵盖了模型中所有存在的用户。标签集表征的是标签空间,其涵盖了模型中存在的所有标签个体,每个标签个体可用短语进行表示,例如"very good",或者利用词语进行表示,例如"Star"。项目集则表征资源空间,其涵盖了模型中存在的所有资源,每个资源都具有唯一编号。

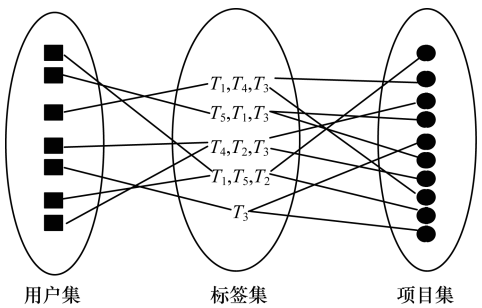

图 3　标签系统的动态关系

1.3　标签偏好评分矩阵计算

基于用户偏好的标签推导过程建立用户与标签项目间的推导预测过程。在资源访问过程中,用户

常表现出 2 类行为:1)对标签的查找、新增和关注行为。2)资源浏览的用户交互过程,例如浏览、点击和收藏等。这 2 类行为,可充分体现出用户对于信息的偏好行为,如果把这种行为的相关特征作为算法预测的权重,那么可实现更好地预测用户偏好。

基于 Sigmoid 函数对标签的质量进行计算,并可基于标签质量获得所采用的相关特征权重。假定项目与对应标签 t 之间的权重值形式为 $\omega(i,t)$,那么可得其计算形式为:

$$\omega(i,t)=\frac{1}{\exp(-m(i,t))} \tag{2}$$

其中,m 是推荐标签的质量,并且满足关系 $m=TF\times IDF$,TF 为词频参数,指的是文档 d 中词条 t 的出现频次,IDF 为文档的逆频率参数,指的是文档频率与词条 t 频率之间存在的反比关系。

可采取多种方式对系统用户之间的交互过程进行标签偏好的预测推导,其中采用评分方法可更好地对用户的偏好进行表达。对此,这里利用数字评分方式表达用户与项目资源之间的偏好关系,这种方法也称为项目评级方法(Item-ratings,IR)。该方法中充分考虑了相关权重在标签偏好推导过程中的作用,则可得:

$$IR(u,t)=\frac{\sum\limits_{i\in M_t}\omega(i,t)\cdot r_{u,i}}{\sum\limits_{i\in M_t}\omega(i,t)} \tag{3}$$

其中,$r_{u,i}$为项目 i 上用户 u 的评分数值;$\omega(i,t)$是标签 t 与项目 i 间存在的关联权重;u 为网络用户;M_t 是标签 t 的所有项目集。但是,上述表示方式并未在分母和分子中考虑未评价项目。对于某些网站或者测试集,无法获得精确预测。但若用户对于某个项目进行收藏等操作表明其偏好,则可认为此类项目是用户偏好的。

在完成用户偏好预测过程后,可进一步采取隐式标签对项目资源用户评分进行预测,具体过程如下:首先,对项目 i 的标签进行用户偏好计算,从而获得其兴趣程度;然后,对标签和对应项目之间的权重 $\omega(i,t)$进行计算。若标签的用户兴趣度计算结果为 $NTP(u,t)$,则可得项目 i 上用户 u 偏好评价值为:

$$IT(u,i)=\sum_{i\in T_i}NTP(u,t)\cdot\omega(i,t) \tag{4}$$

2 概率矩阵分解的推荐算法

2.1 模型与框架

一般意义上的协同过滤推荐过程设计思路如下:若某个项目上具有相似的 2 组用户评分,则其他项目上这 2 组用户的评分也相似,但实际上这种关系的存在具有一定局限性。而协同过滤的信任推荐算法设计思路为:基于相似偏好对用户之间的信任关系进行建立,具体推荐过程如图 4 所示。

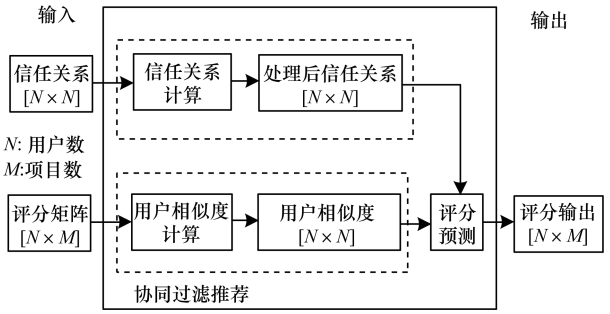

图 4 协同过滤的信任推荐过程

针对不同的信任来源,可将其分成隐性信任和直接信任 2 种来源形式。其中,采用直接信任来源的协同过滤推荐可实现对网络社会属性的充分利用,并基于朋友关系对用户影响力进行定义。

这里所采用的信任主要指的是隐性信任,可基于历史用户评分矩阵进行计算。假定用户偏好越相似,则表征用户之间具有越高的信任值,反之亦然,同时信任程度的影响因素也是多方面的。在进行网络信任模型构建过程中,这里充分利用用户影响力、偏好相似度以及用户专业程度进行影响因子选取和模型建立,所采取的网络信任模型如图 5 所示。

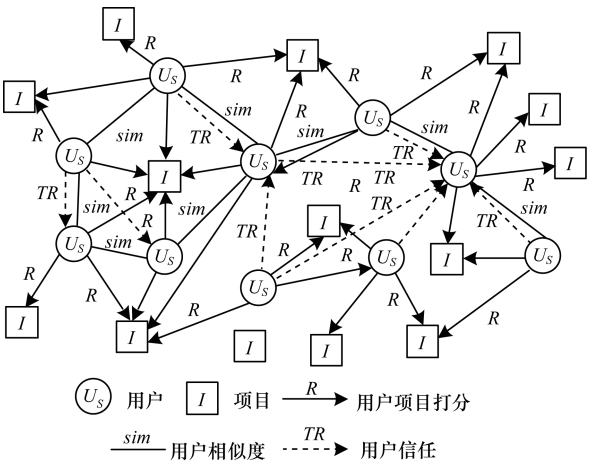

图 5 网络信任模型

在网络的社会属性下,节点之间的关系可表征其存在的实质连接,而信任关系是节点之间关系的一种,可采用用户偏好进行信任信息的获取。对于传统协同推荐过程,一个关键因素是用户相似度。文献[13]利用相似度设定阈值,建立用户信任值:

$$T'_{u,v}=\begin{cases}sim(u,v), sim(u,v)\geqslant\theta_1\\ 0, \text{otherwise}\end{cases} \tag{5}$$

如前所述,用户信任满足非对称特征,其表明用户的信任具有转移性和方向性特征,基于非对称的信任信息,可实现对协同过滤过程的数据稀疏问题缓解。因此,可基于用户之间的影响力关系,实现对用户信任非对称的测度,具体可用 Jaccard 距离进行表征:

$$T''_{u,v}=\frac{IT_u\cap IT_v}{IT_u} \tag{6}$$

基于设定的打分项阈值,可实现用户打分偏差的计算,计算形式为:

$$Tr(u)=\frac{\sum_{i=1}^{M}(R_{ui}-\hat{R}_i)I_{ui}}{|I_u|},|I_u|\geqslant\theta \tag{7}$$

其中,若满足条件 $R_{ui}=0$,则可得 $I_{ui}=0$,否则可得 $I_{ui}=1$。那么最终的信任值计算形式为:

$$T_{u,v}=(\lambda T'_{u,v}+(1-\lambda)T''_{u,v})\cdot Tr(u) \tag{8}$$

2.2　模型分析与计算

在进行模型参数计算过程中,为确保信任评分矩阵具有最大的联合概率 $P(\boldsymbol{R},\boldsymbol{T}|\Lambda)$,从而确保获得最佳近似形式 $u_{1:N}$ 和 $v_{1:M}$。信任评分矩阵应满足如下联合概率条件:

$$P(\boldsymbol{R},\boldsymbol{T}|\Lambda)\propto P(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\tau_R^2)P(\boldsymbol{V}|\boldsymbol{u}_1,\boldsymbol{\Sigma}_1)$$
$$P(\boldsymbol{U}|\boldsymbol{u}_2,\boldsymbol{\Sigma}_2)\times P(\boldsymbol{F}|\boldsymbol{u}_3,\boldsymbol{\Sigma}_3)P(\boldsymbol{T}|\boldsymbol{U},\boldsymbol{F},\tau_T^2) \tag{9}$$

其中,$\boldsymbol{R}$ 为评分矩阵;$\boldsymbol{T}$ 为信任矩阵;$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为特征矩阵。模型的参数集形式为:

$$\Lambda=\{\boldsymbol{u}_1,\boldsymbol{\Sigma}_1,\boldsymbol{u}_2,\boldsymbol{\Sigma}_2,\boldsymbol{u}_3,\boldsymbol{\Sigma}_3,\tau_R^2,\tau_T^2\}$$

那么对于潜在的特征矩阵,网络模型中用户与标签的先验高斯概率分布为:

$$P(\boldsymbol{V}|\boldsymbol{u}_1,\boldsymbol{\Sigma}_1)=\prod_{j=1}^{M}N(\boldsymbol{v}_j|\boldsymbol{u}_1,\boldsymbol{\Sigma}_1) \tag{10}$$

$$P(\boldsymbol{U}|\boldsymbol{u}_2,\boldsymbol{\Sigma}_2)=\prod_{i=1}^{N}N(\boldsymbol{u}_i|\boldsymbol{u}_2,\boldsymbol{\Sigma}_2) \tag{11}$$

$$P(\boldsymbol{F}|\boldsymbol{u}_3,\boldsymbol{\Sigma}_3)=\prod_{s=1}^{N}N(\boldsymbol{f}_s|\boldsymbol{u}_3,\boldsymbol{\Sigma}_3) \tag{12}$$

此外,$P(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\tau_R^2)$ 和 $P(\boldsymbol{T}|\boldsymbol{U},\boldsymbol{F},\tau_T^2)$ 计算形式为:

$$P(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\tau_R^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}P(\boldsymbol{R}_{ij}|g(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j),\tau_R^2)\sigma_{ij} \tag{13}$$

$$P(\boldsymbol{T}|\boldsymbol{U},\boldsymbol{F},\tau_T^2)=\prod_{i=1}^{N}\prod_{s=1}^{N}P(\boldsymbol{T}_{i,s}|g(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{f}_s),\tau_T^2)\sigma_{is} \tag{14}$$

由此可得,信任评分矩阵的最大联合概率形式为:

$$\begin{aligned}&P(\boldsymbol{R},\boldsymbol{T}|\Lambda)\\&=\int_{\boldsymbol{u}_{1:N}}\int_{\boldsymbol{v}_{1:M}}\int_{f_{1:N}}\prod_{j=1}^{M}P(\boldsymbol{v}_j|\boldsymbol{u}_1,\boldsymbol{\Sigma}_1)\prod_{i=1}^{N}P(\boldsymbol{u}_i|\boldsymbol{u}_2,\boldsymbol{\Sigma}_2)\\&\prod_{i=1}^{N}\prod_{j=1}^{M}P(\boldsymbol{R}_{ij}|g(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j),\tau_R^2)\sigma_{ij}\prod_{s=1}^{N}P(F_s|\boldsymbol{u}_3,\boldsymbol{\Sigma}_3)\\&\prod_{i=1}^{N}\prod_{s=1}^{N}P(\boldsymbol{T}_{is}|g(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{f}_s),\tau_T^2)\sigma_{is}d_{u_{1:N}}d_{v_{1:M}}d_{f_{1:N}}\end{aligned} \tag{15}$$

其中,$\boldsymbol{R}$ 代表评分矩阵;$\boldsymbol{T}$ 代表信任矩阵;$\boldsymbol{u}_i$,$\boldsymbol{v}_j$ 和 $\boldsymbol{f}_s$ 代表特征向量;$g(\cdot)$ 是正则化逻辑映射,计算形式为 $g(x)=1/(1+\exp(-x))$,可确保 $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{f}_s$ 与 $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j$ 的取值区间为[0,1];若满足 $\boldsymbol{R}_{ij}\neq\varnothing$,那么可得 $\sigma_{ij}=1$,否则可得 $\sigma_{ij}=0$。类似的,若满足 $\boldsymbol{T}_{is}\neq\varnothing$,那么可得 $\sigma_{uv}=1$,否则可得 $\sigma_{uv}=0$。对于给定的信任矩阵 $\boldsymbol{T}$ 和评分矩阵 $\boldsymbol{R}$,对 TPMDM 模型进行训练的目标是,获得能够使联合概率 $P(\boldsymbol{R},\boldsymbol{T}|\Lambda)$ 取值最大的模型估计参数 Λ。该求解过程可基于期望最大法求解[14],具体过程:

1)采用 E-step 步骤,对潜在变量的后验概率 $P(u_{1:N},v_{1:M},f_{1:N}|\boldsymbol{R},\boldsymbol{T},\Lambda)$ 进行计算。

2)采用 M-step 步骤,利用模型估计参数 Λ。在后验概率 $P(u_{1:N},v_{1:M},f_{1:N}|\boldsymbol{R},\boldsymbol{T},\Lambda)$ 真实值中引入近似取值 $q(u_{1:N},v_{1:M},f_{1:N}|\Lambda')$,该近似取值中 Λ' 为变化参数,且满足 $\Lambda'=\{\boldsymbol{\lambda}_{1i},v_{1i}^2,\boldsymbol{\lambda}_{2j},v_{2j}^2,\boldsymbol{\lambda}_{3s},v_{3s}^2\}$,则可得 q 形式为:

$$\begin{aligned}&q(u_{1:N},v_{1:M},f_{1:N}|\Lambda')=\prod_{i=1}^{N}q(\boldsymbol{u}_i|\boldsymbol{\lambda}_{1i},\mathrm{diag}(v_{1i}^2))\\&\prod_{i=1}^{N}q(\boldsymbol{v}_j|\boldsymbol{\lambda}_{2j},\mathrm{diag}(v_{2j}^2))\prod_{i=1}^{N}q(\boldsymbol{f}_s|\boldsymbol{\lambda}_{3s},\mathrm{diag}(v_{3s}^2))\end{aligned} \tag{16}$$

由此可得,上述联合概率 $P(\boldsymbol{R},\boldsymbol{T}|\Lambda)$ 最大化求解过程,演变为目标 $L(\Lambda,\Lambda')$ 的优化过程。反复利用 M-step 和 E-step 的迭代优化,可实现 Λ,Λ' 取值的不断迭代更新,最终获得满足联合概率 $P(\boldsymbol{R},\boldsymbol{T}|\Lambda)$ 最大化的 Λ,Λ' 取值。在具体的评分预测过程中,可利用 MAP 过程进行估计,且有:

$$\begin{aligned}\{\boldsymbol{u}_i,\boldsymbol{v}_j,\boldsymbol{f}_s\}&=\underset{u_i,v_j,f_s}{\arg\max}(P(u_{1:N},v_{1:M},f_{1:N},\boldsymbol{R},\boldsymbol{T}))\\&\approx\underset{u_i,v_j,f_s}{\arg\max}(u_{1:N},v_{1:M},f_{1:N}|\Lambda')\\&=(\boldsymbol{\lambda}_{1i},\boldsymbol{\lambda}_{2j},\boldsymbol{\lambda}_{3s})\end{aligned} \tag{17}$$

进而可得,最终的评分估值可计算为:

$$\hat{\boldsymbol{R}}_{ij}=\boldsymbol{\lambda}_{1i}^{\mathrm{T}}\boldsymbol{\lambda}_{2j} \tag{18}$$

综上所述,信任概率矩阵分解模型的训练步骤的算法步骤:

步骤 1　生成随机矩阵 $\boldsymbol{\lambda}_{1i},\boldsymbol{\lambda}_{2j},\boldsymbol{\lambda}_{3s}$。

步骤 2　根据文献[10]对于数据的训练过程计算得到参数 $\Lambda=\{\boldsymbol{u}_1,S_1,\boldsymbol{u}_2,S_2,\boldsymbol{u}_3,S_3,t_R^2,t_T^2\}$。

步骤 3　判定信任矩阵 $\boldsymbol{T}$ 和评分矩阵 $\boldsymbol{R}$ 预测误差是否符合设定的条件 $e_1\leqslant epsilon$,$e_2\leqslant epsilon$,在这里选取 $epsilon=0.000\,1$,并且满足 $t\geqslant\min step$。

步骤 4　根据文献[14]中 M-step 的迭代优化方法获得更新后的参数 $\Lambda'=\{\boldsymbol{\lambda}_{1i},v_{1i}^2,\boldsymbol{\lambda}_{2j},v_{2j}^2,\boldsymbol{\lambda}_{3s},v_{3s}^2\}$。

步骤 5　根据文献[14]中 E-step 的迭代优化方法获得更新后的参数 $\Lambda=\{\boldsymbol{u}_1,\boldsymbol{\Sigma}_1,\boldsymbol{u}_2,\boldsymbol{\Sigma}_2,\boldsymbol{u}_3,\boldsymbol{\Sigma}_3,\tau_R^2,\tau_T^2\}$。

步骤 6　根据 $\boldsymbol{\lambda}_{1i}^{\mathrm{T}}\boldsymbol{\lambda}_{2j}$ 得到输出评分矩阵 $\hat{\boldsymbol{R}}_{ij}$。

3　实验结果及分析

3.1　实验设置

为进一步验证所提 TPMDM 模型算法的性能优势,设计如下两组实验对比过程:1)在协同过滤推荐过

程中,阈值 θ 的影响程度;2)选取对比算法,验证所提 TPMDM 模型算法与其他过滤推荐方法的性能优劣。

实验验证过程所使用的测试集有 2 组,分别是 Jester-data 测试集以及 MovieLens 测试集。2 组测试集下载网站为 http://www.grouplens.org/。在 MovieLens 测试集中,共含有 974 组用户,1 743 组电影,100 000 条标签评分信息,其中每组用户最低存有 18 组标签评分记录。因为 Jester-lata 测试集内部的数据存储过于稠密,所以采取随机方式,在其中选择 1 849 组用户对于网络模型中的 116 组新闻的共计 18 963 组标签评分信息,每组用户最低含有 1 组标签评分信息。选取的 2 组测试集的数据信息如表 1所示。

表 1 测试集参数

参数	Jester-lata	MovieLens
规模	1 847 × 100	941 × 1 678
评分区间	-10 ~ 10	1 ~ 5
最低评分记录	1	22
稀疏程度	0.865 4	0.936 8

为确保实验过程的公正性,以及实验对比分析的无偏性,采取 5-fold 交叉验证方法。将测试集分解为训练和测试 2 组集合,分别占整个测试集的 80% 和 20% ,并且训练和测试 2 组集合存在互斥关系,且能够实现对测试集的整体覆盖。

3.2 评价指标

评价指标如下:

1)绝对误差均值指标。该指标利用用户预测评分与真实评分之间的偏差进行评价指标计算,并表征评分预测的精度。绝对误差均值取值越大,表明预测算法精度越差。对于给定的 n 组项目项目,假定用户预测评分集形式为 $p=\{p_1,p_2,\cdots,p_n\}$,而用户真实评分集形式为 $r=\{r_1,r_2,\cdots,r_n\}$,那么可得绝对误差均值指标的计算形式为:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|p_i-r_i| \tag{19}$$

2)覆盖率指标。该指标表征可实现评分预测的项目占比,该指标越大表示预测项目越全面。对于给定的 n 组项目,如果用户评分集预测为 $p=\{p_1,p_2,\cdots,p_n\}$,并且利用 g 表征集合 p 评分数量,那么可得覆盖率指标计算形式为:

$$Coverage=\frac{g}{n} \tag{20}$$

3)精度指标与召回率指标。假定模型中的用户 u 对于测试集的正面评分集是 T_u,则在其中选择 N 个最大评分项,获得用户评分的 *Top-N* 推荐集 R_u。定义项目推荐精度指标,该指标数值越大,表明协同过滤推荐过程的精度越高。召回率指标表征系统的召回率,该指标取值越高,表明数据的协同过滤推荐越全面。这 2 组指标的具体计算形式为:

$$Precision=\frac{1}{|U|}\sum_{u\in U}\frac{|T_u\cap R_u|}{|R_u|} \tag{21}$$

$$Recall=\frac{1}{|U|}\sum_{u\in U}\frac{|T_u\cup R_u|}{|R_u|} \tag{22}$$

3.3 结果分析

3.3.1 阈值实验结果

为获得式(5)中设定的阈值参数 θ 对推荐算法的性能影响,这里设计算法对其进行验证,设定近邻数参数值为 $k=10$,实验数据见图 6。

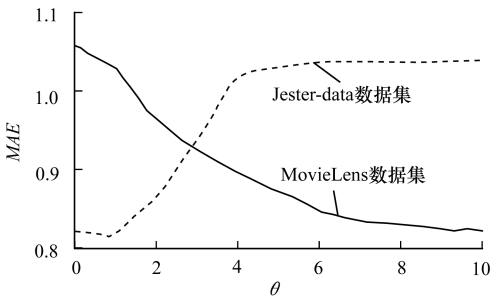

图 6 阈值实验结果

图 6 中给出本文算法在 MovieLens 测试集和 Jester-data 测试集上的绝对误差均值随阈值 θ 变化情况。根据图 6 可知,随着阈值 θ 增大,MovieLens 测试集上获得的 MAE 指标始终处于下降态势,特点是先快速下降,后趋于平稳。而在 Jester-data 测试集上,其 MAE 指标先呈现缓慢下降态势,后逐渐增大,最终趋于平稳。可见本文算法在 MovieLens 测试集和 Jester-data 测试集上的绝对误差均值变化情况不一致,出现这种趋势的主要原因是,Jester-data 测试集在规模上要小于 MovieLens 测试集,因此,在 θ 取值过大时,Jester-data 测试集的用户相似性不可计算,造成协同过滤推荐的精度反而出现下降。而在 MovieLens 测试集上,规模较大时用户相似性不可计算阈值会相应增加,这种情况下的 θ 选取较大值设定有利于协同过滤推荐精度提升。

根据图 6 实验数据可知,对于 Jester-data 测试集,设定阈值 $\theta=1$ 所得协同过滤推荐的精度最高。对于 MovieLens 测试集,设定阈值 $\theta=6$ 所得协同过滤推荐的精度最高。

3.3.2 协同过滤推荐质量对比

为更加充分的验证所提推荐算法性能优势,这里选取 2 个经典的协同过滤推荐算法进行性能对比验证,分别为余弦协同过滤(Cosine Collaborative Filtering, COSCF)和采用皮尔逊相关系数协同过滤(Pearson Correlation Coefficient Collaborative Filtering, PCCCF)算法,此外选取文献[15]提出的新启发式相识度模型协同过滤(New Heuristic Similarity Model Collaborative Filtering, NHSMCF)算法,共 3 种算法进行对比[15],实验结果如图 7、图 8 所示。

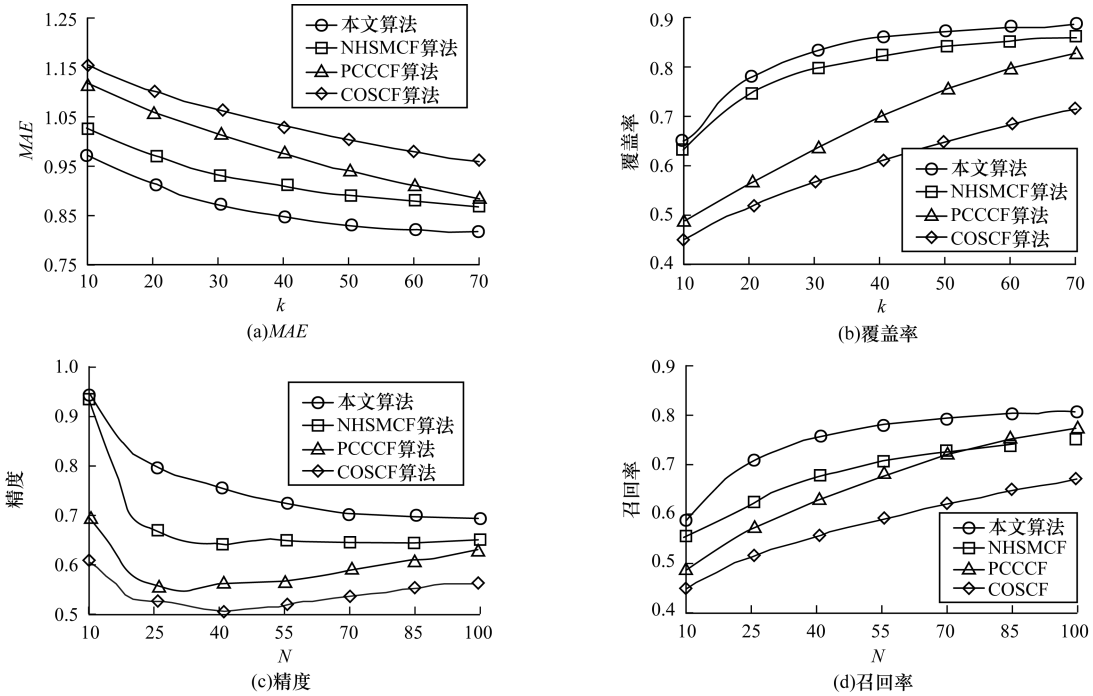

图 7　MovieLens 测试集对比指标

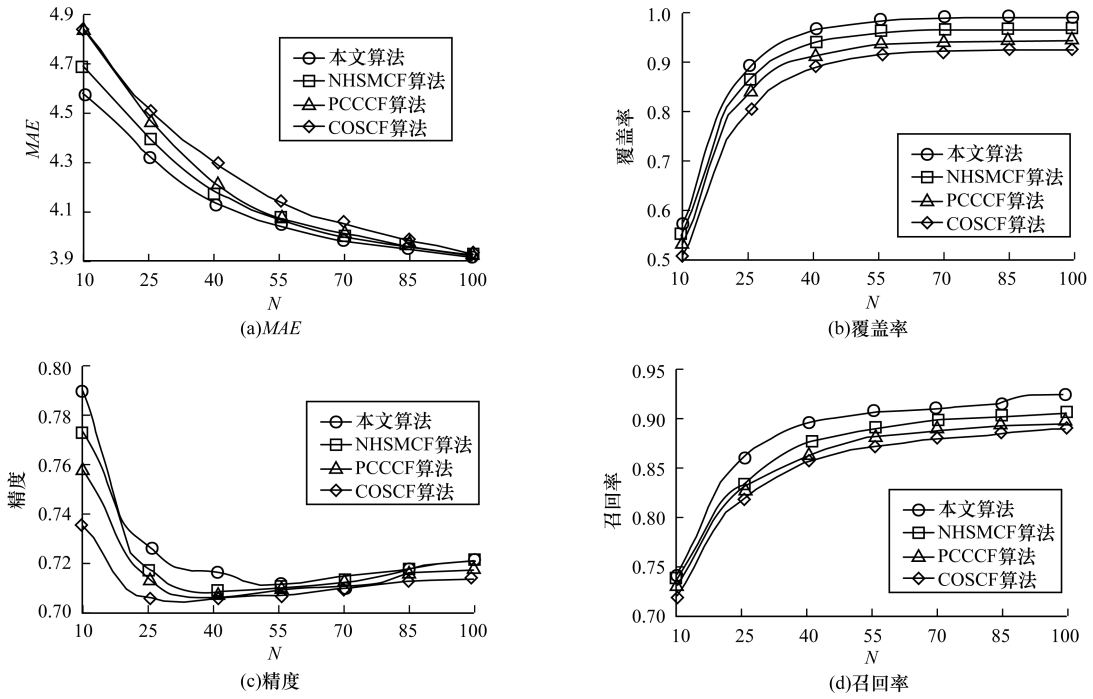

图 8　Jester-data 测试集对比指标

根据图 7 可知,本文算法在选取的 4 种评价指标中,其性能表现均要 COSCF、PCCCF 以及 NHSMCF 3 种算法,且随设定的项目推荐数量 N 以及近邻参数 k 的增大,所提算法的协同过滤推荐质量逐渐提高。其余 3 种对比算法,NHSMCF 算法的推荐质量要高于 COSCF 和 PCCCF 算法,COSCF 算法的推荐质量最差。根据图 8 可知,Jester-data 测试集与 MovieLens 测试集的协同过滤推荐质量对比情况相似,可以看出,在相同条件设置下,所提算法在绝对误差均值指标、覆盖率指标、精度指标与召回率指标上均要优于选取的对比算法。且随设定的项目推荐数量 N 以及近邻参数 k 的增大,所提算法的协同过滤推荐质量逐渐提高。其余 3 种对比算法情况与 MovieLens 测试集实验结果近似,NHSMCF 算法的推荐质量要高于 COSCF 和 PCCCF 算法,COSCF 算法的推荐质量最差。

4 结束语

基于动态标签偏好信任概率矩阵分解模型,本文提出一种改进的协同过滤推荐算法。利用动态关系进行信任评分值计算,构建 TPMDM 模型并采用期望最大法对模型进行求解。实验结果表明,该算法具有较好的推荐质量性能。由于算法是利用评分对信任关系进行构建,因此下一步将就如何更好构建协同过滤推荐的信任模型,以及用户偏好的非信任关系做深入研究。

参考文献

[1] 刘海洋,王志海,黄 丹,等. 基于评分矩阵局部低秩假设的成列协同排名算法[J]. 软件学报,2015,26(11):2981-2993.

[2] YOSHIZAKI S,YOSHITOMI Y,KORO C,et al. Music Recommendation Hybrid System for Improving Recognition Ability Using Collaborative Filtering and Impression Words[J]. Artificial Life and Robotics,2013,18(1):109-116.

[3] ANTOINE B,DAVIDE F,RACHID G,et al. Privacy-preserving Distributed Collaborative Filtering[J]. Computing,2016,98(8):827-846.

[4] MOJTABA S,ISA N K,MOHAMMAD B,et al. Personalized Recommendation of Learning Material Using Sequential Pattern Mining and Attribute Based Collaborative Filtering[J]. Education and Information Technologies,2014,19(4):713-735.

[5] 徐 蕾,杨 成,姜春晓,等. 协同过滤推荐系统中的用户博弈[J]. 计算机学报,2016,39(6):1176-1188.

[6] LI Yanen,ZHAI Chengxiang,CHEN Ye. Exploiting Rich User Information for One-class Collaborative Filtering[J]. Knowledge and Information Systems,2014,38(2):277-301.

[7] 张燕平,张 顺,钱付兰,等. 基于用户声誉的鲁棒协同推荐算法[J]. 自动化学报,2015,41(5):1004-1011.

[8] ADITYA K M,JIANG Xiaoqian,JIHOON K,et al. Detecting Inappropriate Access to Electronic Health Records Using Collaborative Filtering[J]. Machine Learning,2014,95(1):87-101.

[9] MARYAM K N,MAHRIN M. A Systematic Literature Review on the State of Research and Practice of Collaborative Filtering Technique and Implicit Feedback[J]. Artificial Intelligence Review,2016,45(2):167-201.

[10] 王兴茂,张兴明,吴毅涛,等. 基于启发式聚类模型和类别相似度的协同过滤推荐算法[J]. 电子学报,2016,44(7):1708-1713.

[11] YUAN Ting,CHENG Jian,ZHANG Xi,et al. Enriching One-class Collaborative Filtering with Content Information from Social Media[J]. Multimedia Systems,2016,22(1):51-62.

[12] CHEN Jiemin,TANG Feiyi,XIAO Jing,et al. CogTime_RMF: Regularized Matrix Factorization with Drifting Cognition Degree for Collaborative Filtering[J]. Cluster Computing,2016,19(2):821-835.

[13] MAHDI N, BEHROUZ M. Increasing Prediction Accuracy in Collaborative Filtering with Initialized Factor Matrices[J]. The Journal of Supercomputing,2016,72(6):2157-2169.

[14] 张晓瀛,张 洪,唐燕群,等. MIMO-OFDM 系统中基于变分 Bayes EM 算法的联合符号检测与鲁棒 Kalman 信道跟踪[J]. 中国科学(信息科学),2013,43(9):1147-1161.

[15] LIU H F,HU Z,MIAN A,et al. A New User Similarity Model to Improve the Accuracy of Collaborative Filtering[J]. Knowledge-based System,2014,56(3):156-166.

编辑 刘 冰

(上接第 159 页)

参考文献

[1] 林 楠,李翠霞. SVM 在非线性网络流量预测中的应用研究[J]. 计算机仿真,2011,28(5):159-162.

[2] 田中大. 遗传算法优化回声状态网络的网络流量预测[J]. 计算机研究与发展,2015,52(5):1137-1145.

[3] 熊 伟. 基于小波的网络流量异常协同相变检测[J]. 计算机应用,2012,32(8):2271-2274.

[4] 陈 静,刘 渊. 融合模拟退火算法的神经网络流量预测研究[J]. 计算机工程与设计,2011,32(6):2138-2145.

[5] 柏 骏,夏靖波,赵小欢. 一种基于 EMD 和 RVM 的自相似网络流量预测模型[J]. 计算机科学,2015,42(1):122-125.

[6] 高 波. 基于 EMD 及 ARMA 的自相似网络流量预测[J]. 通信学报,2011,32(4):47-56.

[7] 袁小坊,陈楠楠,王东城. 城域网应用层流量预测模型[J]. 计算机研究与发展,2009,46(3):434-442.

[8] 董春玲. 一种结合 DWT 和 FARIMA 的网络拥塞控制机制[J]. 小型微型计算机系统,2011,32(5):931-934.

[9] HERNANDEZ L, BALADRON C, AGUIAR J M, et al. Artificial Neural Networks for Short-term Load Forecasting in Microgrids Environment[J]. Energy,2014,75(1):252-264.

[10] 杜 涛. 基于 BP 神经网络技术的网络流量预测模型[J]. 网络安全技术与应用,2016,36(7):55-57.

[11] 李小航,刘 渊,刘元珍. 基于小波多尺度分析的网络流量组合预测方法研究[J]. 微电子学与计算机,2008,25(1):130-133.

[12] 冯兴杰,潘文欣,卢 楠. 基于小波包的 RBF 神经网络网络流量混沌预测[J]. 计算机工程与设计,2012,33(5):1681-1686.

[13] 马 力,张高明,苟娟迎. 一种基于小波变换的校园网流量预测方法研究[J]. 计算机科学,2012,39(z2):63-73.

[14] SEYEDALI M. Moth-flame Optimization Algorithm[J]. Knowledge-based System,2015,89(c):228-249.

[15] MENG A,CHEN Y,YIN H,et al. Crisscross Optimization Algorithm and Its Application[J]. Knowledge-based System,2014,67(4):218-229.

[16] 吴伟民,林水明,林志毅. 一种基于混沌不透明谓词的压扁控制流算法[J]. 计算机科学,2015,42(5):178-182.

[17] 徐 斐,谢洲烨,沈 伟,等. 基于神经网络的分布式雷达抗干扰效能评估方法[J]. 现代雷达,2015,37(7):4-7.

[18] LIU H, TIAN H, CHEN C, et al. An Experimental Investigation of Two Wavelet-MLP Hybrid Frameworks for Wind Speed Prediction Using GA and PSO Optimization[J]. International Journal of Electrical Power & Energy Systems,2013,52(1):161-173.

编辑 刘 冰