文章编号:1674-2974(2015)10-0107-07

基于 PMF 进行潜在特征因子分解的标签推荐*

刘胜宗[1,3],樊晓平[1,3],廖志芳[2†],吴言凤[2]
(1.中南大学 信息科学与工程学院,湖南 长沙 410075;2.中南大学 软件学院,湖南 长沙 410075;
3.湖南财政经济学院 网络化系统研究所,湖南 长沙 410205)

摘 要:现有社会标签推荐技术存在数据稀疏、时间复杂度高以及可解释性低等问题,鉴于此,提出基于概率矩阵分解(PMF)进行潜在特征因子联合分解的标签推荐算法(TagRec-UPMF),它结合用户、资源及标签3方面的潜在特征,联合构建对应的概率形式的潜在特征向量,然后根据它们两两之间的特征向量内积进行线性组合,从而产生 Top-N 推荐.该算法解决了数据规模大且稀疏情况下的精度问题,算法的线性复杂度使得其可用于大规模数据.实验结果表明,相比于 TagRec-CF,PITF, TTD,Tucker,NMF 等算法,本文算法既提高了推荐的准确率,又降低了时间损耗.与 PITF 算法相比较,准确率得到了提高,而处理时间相差不明显;与 TTD 算法相比较,在准确率相差不明显的情况下,大大降低了时间损耗.因此,本文的 TagRec-UPMF 算法相比其他算法表现出了一定的优势.

关键词:协同过滤;潜在特征因子;标签推荐;推荐系统;概率矩阵分解
中图分类号:TN911.23 **文献标识码**:A
DOI:10.16339/j.cnki.hdxbzkb.2015.10.018

A Tag Recommending Algorithm with Latent Feature Factor Jointly Factorizing Based on PMF

LIU Sheng-zong[1,3], FAN Xiao-ping[1,3], LIAO Zhi-fang[2†], WU Yan-feng[2]
(1. School of Information Science and Engineering, Central South Univ, Changsha, Hunan 410075, China;
2. School of Software, Central South Univ, Changsha, Hunan 410075, China;
3. Laboratory of Networked Systems, Hunan Univ of Finance and Economics, Changsha, Hunan 410205, China)

Abstract: The existing social tag recommending technology has the problems of data sparsity, high time complexity and low interpretability. To solve these problems, this paper proposed a tag recommending approach called TagRec-UPMF, which jointly factorizes the latent feature factor based on PMF. The approach jointly builds the corresponding feature vector in the form of probability, combining latent features of the three different facets of users, resources and tags, and then produces the top-N recommendation according to the linear combination of the inner products between the feature vectors of each pair. The proposed algorithm improves its accuracy in the case of the large size and sparse data, and it can be used

* 收稿日期:2014-07-19
基金项目:国家科技支撑计划资助项目(2012BAH08B00);国家自然科学基金资助项目(61073105),National Natural Science Foundation of China(61073105);国家自然科学青年基金资助项目(61202095),Youth Fund of National Natural Science Foundation Projects (61202095);计算机应用技术湖南省"十二五"重点建设学科资助项目;信息技术与信息安全湖南省普通高等学校重点实验室资助项目
作者简介:刘胜宗(1986—),男,湖南邵阳人,中南大学博士研究生
† 通讯联系人,E-mail:liaozf415@126.com

for large-scale data due to the linear complexity. Experimental results show that our method has higher accuracy and lower time consuming than TagRec-CF, and Tucker, NMF, etc. Meanwhile, the proposed method has better precision than PITF algorithm when their complexity is of little difference. And our method shows lower complexity compared with TTD algorithm while their precision are nearly the same.

Key words: collaborative filtering; latent feature factor; tag recommender; recommendation system; probabilistic matrix factorization

作为 Web2.0 的重要特征,社会标签系统允许用户对系统资源利用个性化标签进行标注,从而使具有相同兴趣偏好的用户相互推荐及共享资源[1].国内外知名社会标签系统有音乐类标签系统 last.fm[2]、图片类标签系统 flickr[3]、电影类标签系统 movielens[4]、书签和出版物信息共享系统 bibsonomy[5]等.这些网站采用社会标签整合各类资源,这有助于用户组织、浏览和搜索自己感兴趣的资源,也能够更好地帮助用户之间进行沟通及共享,而标签推荐系统可将用户感兴趣的标签推荐给使用同一资源的用户[6].

标签推荐系统基于用户以往的标注行为进行标签推荐,这种推荐同时依赖于用户和资源[7].目前广泛应用的协同过滤推荐[8](CF)为目标用户寻找有相似标注行为的其他用户(近邻),并将近邻在目标资源上标注过的其他标签推荐给目标用户,该技术简单和实用,但也面临着冷启动和数据稀疏问题[6].基于此,研究者尝试从其他角度去研究新的推荐策略及方法,目前,大部分关于标签推荐的研究集中在因子分解方面,比较典型的有非负矩阵分解(NMF)[9],奇异值分解(SVD)[10],高阶奇异值分解(HOSVD)[6],Tucker 张量分解[8],PITF 张量分解[1]以及 TTD 张量分解[6],这些方法在解决数据稀疏性和缺失值带来的问题上取得了较好的效果.但这些分解技术仅考虑了标注关系,并未考虑用户的评分偏好关系,由于用户选择标签进行标注的过程中同时受自身对资源和标签的兴趣偏好影响.另外,不同用户对标签或资源的兴趣偏好侧重面不一样[11],标签和资源是受某些基本的、潜在的特征支配,用户的偏好则是由用户对这些潜在特征喜好程度的加权综合,用户的标注行为除受本身偏好的影响之外,同样还受到标签和资源的潜在特征结构的影响[8].这体现出一种"资源-标签"的双重概率关系[12],这种关系同样存在于"用户-资源"、"用户-标签"情形中.为了解决上述问题,本文提出一种新的标签推荐方法(TagRec-UPMF),该方法采用概率矩阵分解技术进行潜在特征因子联合分解,然后通过潜在特征向量之间相互组合完成推荐.

1 问题定义

社会标签系统可形式化定义为 $F:=(\mathrm{US},\mathrm{TS},\mathrm{IS},\mathrm{RS})$,其中 US 为 User 集合,TS 为 Tag 集合,IS 为 Item 集合,RS 为 User,Item 和 Tag 之间的关系集合,其中 $\mathrm{RS}\in\mathrm{TS}\times\mathrm{US}\times\mathrm{IS}$[6].标签推荐是在用户访问的资源上推荐与资源相关的标签.符号标记如表 1 所示.

表 1 符号标记表

Tab. 1 Definition table of symbol

符号标记	解释说明
$\mathrm{US}=\{u_1,u_2,\cdots,u_m\}$	用户集合,共 m 个用户
$\mathrm{IS}=\{i_1,i_2,\cdots,i_n\}$	资源集合,共 n 个资源
$\mathrm{TS}=\{t_1,t_2,\cdots,t_o\}$	标签集合,共 o 个标签
$\boldsymbol{U}\in\mathbf{R}^{l\times m}$	用户潜在特征矩阵
$\boldsymbol{V}\in\mathbf{R}^{l\times n}$	资源潜在特征矩阵
$\boldsymbol{W}\in\mathbf{R}^{l\times o}$	标签潜在特征矩阵
$l\in\mathbf{R}$	潜在特征空间维数
$\boldsymbol{B}=\{b_{ui}\},\boldsymbol{B}\in\mathbf{R}^{m\times n}$	用户-资源关系矩阵
$\boldsymbol{C}=\{b_{ut}\},\boldsymbol{C}\in\mathbf{R}^{m\times o}$	用户-标签关系矩阵
$\boldsymbol{A}=\{a_{it}\},\boldsymbol{A}\in\mathbf{R}^{n\times o}$	资源-标签关系矩阵

一般地,用户对标签的认可程度、用户对资源的兴趣程度和资源与标签的关联程度分别表示成用户-标签认可关系矩阵 $\boldsymbol{C}$、用户-资源兴趣矩阵 $\boldsymbol{B}$ 和资源-标签关联度矩阵 $\boldsymbol{A}$.l 表示潜在特征空间的维数.用户对标签的认可程度由用户潜在特征向量和标签潜在特征向量的内积得到,用户对资源的兴趣程度由用户潜在特征向量和资源潜在特征向量的内积得到,资源与标签的关联度由资源潜在特征向量和标签潜在特征向量的内积得到.

设用户 u 访问资源 i 时,选择标签 t 的概率表示为 $y_{u,i,t}$,那么

$$y_{u,i,t}=f(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i,\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t,\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t). \tag{1}$$

式中:$\boldsymbol{U}_u$ 为用户 u 的潜在特征向量;$\boldsymbol{V}_i$ 为资源 i 的

潜在特征向量;W_t 为标签 t 的潜在特征向量;$U_u^{\mathrm{T}}V_i$,$U_u^{\mathrm{T}}W_t$,$V_i^{\mathrm{T}}W_t$ 分别用于计算用户 u 对资源 i 的感兴趣程度、用户 u 对标签 t 的认可程度以及标签 t 与资源 i 的关联程度;$f(\cdot)$参数为 $U_u^{\mathrm{T}}V_i$,$U_u^{\mathrm{T}}W_t$,$V_i^{\mathrm{T}}W_t$ 的函数;$y_{u,i,t}$又称为给定用户 u 和资源 i 情况下的标签 t 的推荐概率.

当用户 u 在访问资源 i 时,标签的 Top-N 推荐列表可以定义如下[6]:

$$\mathrm{Top}(u,i,N):=\underset{t\in T}{\overset{N}{\mathrm{argmax}}}(y_{u,i,t}). \tag{2}$$

式中:N 为推荐列表的长度.

2　基于 UPMF 的标签推荐模型

本文提出基于联合概率矩阵分解(UPMF)的标签推荐算法 TagRec-UPMF,算法包含 3 个部分:

1)求解潜在特征向量. 首先根据训练数据集计算实体间的关系矩阵,然后根据分解算法通过梯度下降方法,以最大化联合的后验概率为目标函数,学习得到用户潜在特征向量、资源潜在特征向量以及标签潜在特征向量.

2)根据公式(3)对给定的用户和资源计算标签集中各标签的推荐概率.

$$y_{u,i,t}=\beta U_u^{\mathrm{T}}V_i+\gamma U_u^{\mathrm{T}}W_t+\delta V_i^{\mathrm{T}}W_t,$$
$$\text{s.t.}\quad \beta+\gamma+\delta=1. \tag{3}$$

3)根据 Top-N 推荐规则,选取推荐概率排名前 N 的标签推荐给用户.

2.1　实体间关系矩阵的计算

1)用户-资源关系矩阵. 用户-资源关系矩阵 B 表示 m 个用户对 n 个资源的兴趣对应关系. B 中元素 b_{ui} 表示用户 u 对资源 i 感兴趣的程度.

$$b_{ui}=\alpha g(h_{ui})+(1-\alpha)g(r_{ui}). \tag{4}$$

式中:h_{ui} 为资源 i 被用户 u 标注的次数;r_{ui} 为用户 u 对资源 i 的评分;$g(\cdot)$为 logistic 函数,用于归一化;α 为平衡因子,取值为[0,1].

2)用户-标签关系矩阵. 用户-标签关系矩阵 C 表示 m 个用户对 o 个标签的偏好对应关系. C 中每个元素 c_{ut} 表示用户 u 对标签 t 的偏好或者认知程度.

$$c_{ut}=g(\lambda_{ut}). \tag{5}$$

式中:λ_{ut} 为用户 u 使用标签 t 的次数.

3)资源-标签关系矩阵. 资源-标签关系矩阵 A 表示 n 个资源和 o 个标签的关联度关系. A 中元素 a_{it} 表示资源 i 和标签 t 之间的关联程度,通常认为,在资源 i 上标注标签 t 的次数越多,表示有越多的用户认为标签 t 和资源 i 的关联度大. a_{it} 由公式(6)计算得到:

$$a_{it}=g(\tau_{it}). \tag{6}$$

式中:τ_{it} 为资源 i 上标注标签 t 的次数.

2.2　概率矩阵联合分解

TagRec-UPMF 标签推荐模型的概率图如图 1 所示.

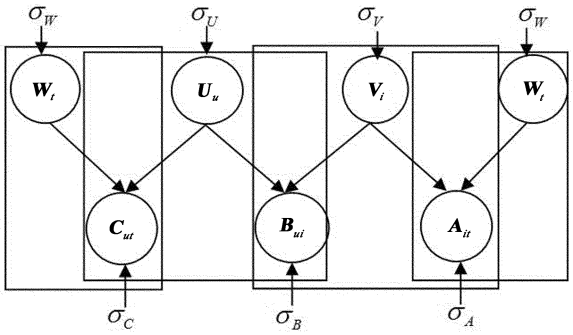

图 1　TagRec-UPMF 的概率图模型

Fig. 1　Probabilistic graphical model of TagRec-UPMF

其中,用户潜在特征向量 U_u 由用户-标签关系信息和用户-资源关系信息共享;资源潜在特征向量 V_i 则由用户-资源关系信息和资源-标签关系信息共享;标签潜在特征向量 W_t 由用户-标签关系信息和资源-标签关系信息共享.

概率矩阵分解模型中,首先假设潜在特征向量 U_u,V_i,W_t 的先验服从均值为 0 的高斯分布,即

$$p(U\mid\sigma_U^2)=\prod_{u=1}^{m}N(U_u\mid 0,\sigma_U^2I); \tag{7}$$

$$p(V\mid\sigma_V^2)=\prod_{i=1}^{n}N(V_i\mid 0,\sigma_V^2I); \tag{8}$$

$$p(W\mid\sigma_W^2)=\prod_{i=1}^{o}N(W_t\mid 0,\sigma_W^2I). \tag{9}$$

在给定用户 u,资源 i 的潜在特征向量(维数为 l)U_u,V_i 后,用户 u 对 i 的感兴趣程度 b_{ui} 满足均值为 $g(U_u^{\mathrm{T}}V_i)$,方差为 σ_B^2 的高斯分布并相互独立,因此 B 的条件概率分布为:

$$p(B\mid U,V,\sigma_B^2)=\prod_{u=1}^{m}\prod_{i=1}^{n}[N(b_{ui}\mid g(U_u^{\mathrm{T}}V_i),\sigma_B^2)]^{I_{ui}^B}. \tag{10}$$

式中:I_{ui}^B 为指示函数,当用户 u 访问或标注过资源 i 时,其值为 1,否则为 0;$g(\cdot)$为 logistic 函数,用于将 $U_u^{\mathrm{T}}V$ 归一化.

用户 u 对标签 t 的兴趣程度 c_{ut} 满足均值为 $g(U_u^{\mathrm{T}}W_t)$ 方差为 σ_C^2 的高斯分布且相互独立,那么 C

的条件概率分布如下:

$$p(\boldsymbol{C}\mid\boldsymbol{U},\boldsymbol{W},\sigma_C^2)=\prod_{u=1}^{m}\prod_{t=1}^{o}\left[N(c_{ut}\mid g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t),\sigma_C^2)\right]^{I_{ut}^C}. \tag{11}$$

其中,当用户 u 使用过标签 t 进行标注时,I_{ut}^C 为 1,否则为 0.

若资源 i 和标签 t 的关联度 a_{it} 满足均值为 $g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t)$,方差为 σ_A^2 的高斯分布且相互独立时,$\boldsymbol{A}$ 的条件概率分布为:

$$p(\boldsymbol{A}\mid\boldsymbol{V},\boldsymbol{W},\sigma_A^2)=\prod_{i=1}^{n}\prod_{t=1}^{o}\left[N(a_{it}\mid g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t),\sigma_A^2)\right]^{I_{it}^A}. \tag{12}$$

其中,当资源 i 和标签 t 有关联时,I_{it}^A 值为 1,否则为 0.

由图 1 可以推导出 $\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$ 的后验分布函数,该分布函数的自然对数形式如公式(13)所示.

公式(13)中,$\mathbb{C}$ 是不依赖于参数的常量. 在概率矩阵分解模型中,需要最大化公式(13),这是一个无约束情况下的优化问题,该问题的求解等价于最小化公式(14).

$$\begin{aligned}\ln p(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}\mid\boldsymbol{B},\boldsymbol{C},\boldsymbol{A},\sigma_W^2,\sigma_V^2,\sigma_U^2,\sigma_A^2,\sigma_C^2,\sigma_B^2)=&\\ -\frac{1}{2\sigma_B^2}\sum_{u=1}^{m}\sum_{i=1}^{n}I_{ui}^B(b_{ui}-g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i))^2-&\\ \frac{1}{2\sigma_C^2}\sum_{u=1}^{m}\sum_{t=1}^{o}I_{ut}^C(c_{ut}-g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t))^2-&\\ \frac{1}{2\sigma_A^2}\sum_{i=1}^{n}\sum_{t=1}^{o}I_{it}^A(a_{it}-g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t))^2-&\\ \frac{1}{2\sigma_U^2}\sum_{u=1}^{m}\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{U}_u-\frac{1}{2\sigma_V^2}\sum_{i=1}^{n}\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{V}_i-\frac{1}{2\sigma_W^2}\sum_{t=1}^{o}\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{W}_t-&\\ \sum_{u=1}^{m}\sum_{i=1}^{n}I_{ui}^B\ln\sigma_B-\sum_{u=1}^{m}\sum_{t=1}^{o}I_{ut}^C\ln\sigma_C-&\\ \sum_{i=1}^{n}\sum_{t=1}^{o}I_{it}^A\ln\sigma_A-l\sum_{u=1}^{m}\ln\sigma_U-l\sum_{i=1}^{n}\ln\sigma_V-&\\ l\sum_{t=1}^{o}\ln\sigma_W+\mathbb{C};\end{aligned} \tag{13}$$

$$\begin{aligned}\Omega(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{B},\boldsymbol{C},\boldsymbol{A})=&\frac{1}{2}\sum_{u=1}^{m}\sum_{i=1}^{n}I_{ui}^B(b_{ui}-\\ &g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i))^2+\frac{\lambda_C}{2}\sum_{u=1}^{m}\sum_{t=1}^{o}I_{ut}^C(c_{ut}-g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t))^2+\\ &\frac{\lambda_A}{2}\sum_{i=1}^{n}\sum_{t=1}^{o}I_{it}^A(a_{it}-g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t))^2+\\ &\frac{\lambda_U}{2}\|\boldsymbol{U}\|_F^2+\frac{\lambda_V}{2}\|\boldsymbol{V}\|_F^2+\frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2;\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial\Omega}{\partial\boldsymbol{U}_u}=&\sum_{i=1}^{n}I_{ui}^B(g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i)-b_{ui})g'(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i)\boldsymbol{V}_i+\\ &\lambda_C\sum_{t=1}^{o}I_{ut}^C(g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t)-c_{ut})g'(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_t)\boldsymbol{W}_t+\lambda_U\boldsymbol{U}_u;\end{aligned} \tag{15}$$

$$\begin{aligned}\frac{\partial\Omega}{\partial\boldsymbol{V}_i}=&\sum_{u=1}^{m}I_{ui}^B(g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i)-b_{ui})g'(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i)\boldsymbol{U}_u+\\ &\lambda_A\sum_{t=1}^{o}I_{it}^A(g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t)-a_{it})g'(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t)W_t+\lambda_V\boldsymbol{V}_i;\end{aligned} \tag{16}$$

$$\begin{aligned}\frac{\partial\Omega}{\partial\boldsymbol{W}_t}=&\lambda_C\sum_{u=1}^{m}I_{ut}^C(g(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_i)-c_{ut})g'(\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{W}_i)\boldsymbol{U}_u+\\ &\lambda_A\sum_{i=1}^{n}I_{it}^A(g(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t)-a_{it})g'(\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{W}_t)\boldsymbol{V}_i+\lambda_W\boldsymbol{W}_t.\end{aligned} \tag{17}$$

公式(14)中:$\lambda_C=\frac{\sigma_B^2}{\sigma_C^2}$;$\lambda_A=\frac{\sigma_B^2}{\sigma_A^2}$;$\lambda_U=\frac{\sigma_B^2}{\sigma_U^2}$;$\lambda_V=\frac{\sigma_B^2}{\sigma_V^2}$;$\lambda_W=\frac{\sigma_B^2}{\sigma_W^2}$;$\|\cdot\|_F^2$ 表示 F 范数. 公式(14)的局部最小值采用梯度下降法进行求解,参数 $\boldsymbol{U}_u,\boldsymbol{V}_i,\boldsymbol{W}_t$ 的梯度下降更新公式分别为公式(15)—式(17).

2.3 算法复杂度分析

在梯度下降法中,算法的时间开销主要取决于目标函数 Ω 及其相应的梯度下降更新公式. 在标签标注数据和用户评分数据中,存在大量的缺失值,这导致 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ 矩阵很稀疏,容易得出公式(14)目标函数的计算时间复杂度为 $O(\rho_B l+\rho_C l+\rho_A l)$,其中 ρ_A,ρ_B,ρ_C 分别表示 3 个实体关系矩阵 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ 的非零元素数目. 同理,梯度下降公式(15)—(17)的计算复杂度分别为 $O(\rho_B l+\rho_C l)$,$O(\rho_B l+\rho_A l)$,$O(\rho_C l+\rho_A l)$. 所以,算法的一步迭代过程中的计算复杂度为 $O(\rho_B l+\rho_C l+\rho_A l)$,这表示算法的时间复杂度随 3 个关系矩阵中观测数据数量呈正线性关系,意味着该算法可应用于大规模的数据.

3 实验结果及分析

3.1 实验设计

3.1.1 数据集

本文选取目前标签推荐研究常用的 2011-10M 版 movielens 数据集,该数据集包含了 2 113 个用户,10 197 部电影以及 13 222 个标签.

3.1.2 算法性能评价指标

目前衡量推荐算法优劣需要同时考虑准确率和召回率,而准确率和召回率[12]指标往往是负相关的,因此为了综合考虑算法的性能,本文选用 F_1 指

标[12]来衡量算法的性能,F_1指标定义见公式(18),其中 Precision 表示准确率,Recall 表示召回率,其计算方法可参考文献[12]. F_1越高,算法的性能越好.

$$F_1 = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

3.1.3　实验设计

为了检验 TagRec-UPMF 算法的推荐效果,本文需要通过实验解决以下几个方面的问题:1)潜在特征向量的维度 l 对推荐性能的影响;2)平衡因子 α 对推荐结果的影响;3)参数 λ_A 和 λ_C 对推荐结果的影响;4) TagRec-UPMF 算法与现有经典标签推荐算法的准确度及时间效率比较.

实验前,为了比较不同数据规模和稀疏情况下算法的效果,分别从实验数据中抽取 90%,70%,50%,30%作为训练集,其余作为测试集进行实验.

实验过程中,通过对训练集尝试不同的参数值,进而在测试集上得到 F_1指标值. 经反复测试得出参数设为 $\alpha = 0.4$, $\beta = \gamma = \delta = 1/3$, $\lambda_C = 1$, $\lambda_A = 0.6$, $\lambda_U = \lambda_V = \lambda_W = 0.05$ 时,算法的效果最优. 在后续的实验中,若无特别说明,这些参数均设为最优值. 同时实验中,Top-N 推荐中取 $N=10$.

3.2　实验分析

3.2.1　参数 l 对推荐性能的影响

该实验用于检测潜在特征向量的维数 l 对推荐算法性能的影响. 图 2 为 l 对算法准确率的影响,图 3 为 l 对算法时间效率的影响. 从图 2 可以看出,随着特征向量维数的增加,推荐准确率慢慢提高,这说明增加潜在特征向量的维数可以提高矩阵分解算法的准确性,而当 $l>15$ 时,精度增加的趋势变缓. 由图 3 可以看出,随着 l 的增大,算法耗费的时间也成正比的增大. 因此出于准确率和时间损耗的平衡考虑,选择 $l=15$.

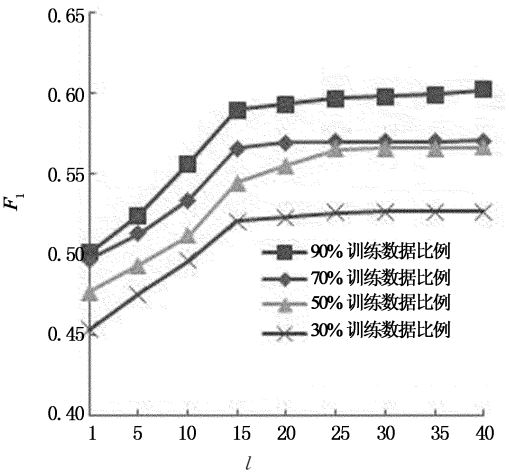

图 2　l 对算法准确率的影响

Fig. 2　Influence on accuracy of l

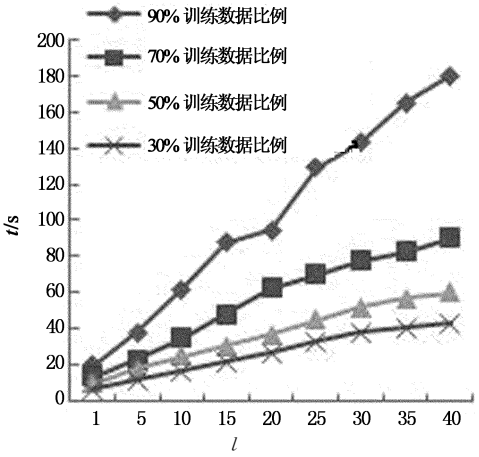

图 3　l 对算法时间消耗的影响

Fig. 3　Influence on complexity of l

3.2.2　α 对推荐准确率的影响

在式(4)中,利用参数 α 来调节资源被标注次数和资源评分在用户对资源兴趣程度中的权重比例,从而影响推荐准确率. 实验结果如图 4 所示. 由图 4 可以看出,α 值处于 0.3 到 0.5 之间时,F_1的值由上升转变为下降趋势,这就意味着在这 2 个值之间存在一个可以使得 F_1最优的 α 值. 本文将 α 值选取为 0.4. 这说明利用资源被标注次数和资源评分的加权组合来表示用户对资源兴趣程度时的效果略好于这两者单独表示的情况.

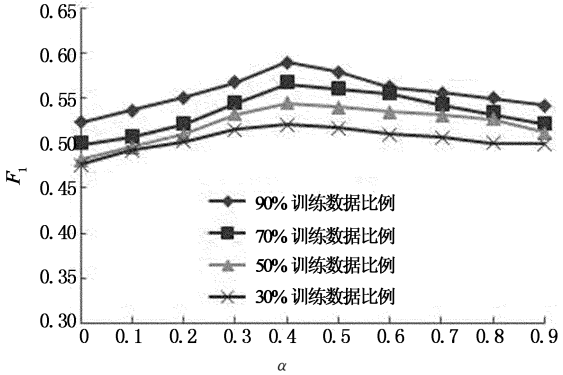

图 4　平衡因子 α 对算法准确率的影响

Fig. 4　Influence on accuracy of α

3.2.3　参数 λ_A 和 λ_C 对推荐结果的影响

概率矩阵联合分解模型有 5 个参数,分别为 λ_A, λ_C, λ_U, λ_V, λ_W,在这部分实验中,主要讨论 λ_A 和 λ_C 的影响,而其他 3 个参数为了简单起见设置为相同的值,并通过交叉验证(cross-validation)的

方式获取这 3 个参数的最优值,即 $\lambda_U = \lambda_V = \lambda_W = 0.05$. TagRec-UPMF 算法中 λ_A 决定了资源-标签关系矩阵对算法的影响权重,而 λ_C 决定了用户-标签关系矩阵对算法的影响权重. 当这两者同时设为 0 时,表示算法在进行推荐时,仅考虑用户-资源关系矩阵,而当 λ_A 或 λ_C 设为 $+\infty$时,则意味着仅利用资源-标签关系矩阵或者用户-标签关系矩阵. 实验结果如图 5 所示,图中显示了在 λ_A 和 λ_C 的不同取值时的算法准确率. 当 $\lambda_A = 1$, $\lambda_C = 0.6$ 时,TagRec-UPMF 算法的准确率最高. 这表明这两个参数相互约束,而用户-标签关系矩阵的影响更显著. 这是因为面向用户推荐标签时,资源和标签之间的相似关系受语义影响较大(多义或同义),而用户和标签之间的关系虽然受用户的主观影响,但依然反映了用户对标签的特殊偏好,因此在推荐过程中需要考虑这两种关系的权衡,也应更多地考虑用户对标签的个性化因素.

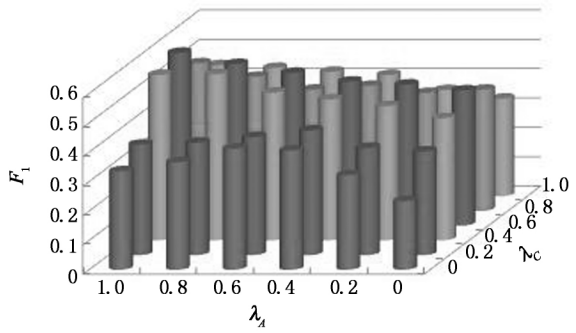

图 5 参数 λ_A 和 λ_C 对算法准确度的影响

Fig. 5 Influence on accuracy of λ_A and λ_C

3.2.4 推荐算法的性能比较

该部分实验是将 TagRec-UPMF 算法和目前常见的部分经典算法从准确率和时间消耗两个方面进行比较,选用的参照算法包括基于协同过滤的标签推荐(TagRec-CF)、基于 Tucker 分解的标签推荐、非负矩阵分解标签推荐算法(NMF)、基于三部图张量分解标签推荐算法(TTD)以及 PITF 算法.

表 2 是在不同训练数据集规模时各算法的 F_1 值(10 次实验结果取平均值). 由表 1 可以看出,在训练数据集比例较小(<50%)时,TagRec-UPMF 算法准确度相对其他算法而言均有提升,当比例较大时,TagRec-UPMF 算法比 TTD 算法的准确度略低,而相比其他算法依然高出 7%~13%,其中 TagRec-CF 算法的准确度受数据稀疏影响最大,准确率最低,实验结果呈现这种现象的原因是 Tucker,NMF,PITF 算法未考虑用户对资源的评分,影响了准确度,而 TTD 算法虽然没考虑评分,但它不仅仅考虑实体间的直接关系,还考虑了两两实体因为第三方实体而产生的间接关系,虽然提高了准确性,但其时间损耗高,在实际应用中并不实用.

表 3 为时间消耗统计情况,其中时间消耗最大的是 Tucker 算法,其次是 TTD 算法,而 PITF 和本文的 TagRec-UPMF 时间消耗最小,PITF 算法的时间消耗略低于 TagRec-UPMF 算法,这是由于 PITF 算法没有考虑评分数据,因此在时间性能上略为占优,但在时间复杂度上,这两者方法依然同为线性级别. 因此,比较各算法在准确率和时间消耗指标上的综合情况,本文的 TagRec-UPMF 算法相比其他算法而言表现出了一定的优势.

表 2 TagRec-UPMF 算法与其他参照算法的准确率比较

Tab. 2 Accuracy comparison between TagRec-UPMF and other reference algorithms

训练数据集比例/%	F_1					
	TagRec-CF	Tucker	NMF	PITF	TTD	TagRec-UPMF
90	0.456 3	0.476 9	0.516 9	0.505 9	0.596 7	0.589 6
70	0.359 6	0.456 1	0.491 2	0.472 7	0.571 3	0.565 9
50	0.156 6	0.432 3	0.468 5	0.456 9	0.498 8	0.544 4
30	0.086 9	0.410 1	0.453 6	0.443 2	0.482 3	0.521 1

表 3 TagRec-UPMF 算法与其他参照算法的时间消耗比较

Tab. 3 Time consuming between TagRec-UPMF and other reference algorithms

训练数据集比例/%	运行时间/min					
	TagRec-CF	Tucker	NMF	PITF	TTD	TagRec-UPMF
90	165	378	243	87	356	88
70	99	201	120	47	183	49
50	65	128	53	27	129	31
30	50	96	34	21	98	22

4 总 结

在社会标签推荐系统中,由于数据非常稀疏,加上现有的标签推荐算法并未充分利用标签标注系统中的相关信息,因此精度不高,而矩阵、张量分解等技术用一种降维的方法表示稀疏数据,缓解了数据稀疏带来的精度问题.本文基于概率矩阵分解,将用户、标签、资源三方面的潜在特征因子进行联合分解,并将求得的特征向量两两之间的内积进行线性加权并产生推荐.在实验过程中讨论了 TagRec-UPMF 算法中各参数对结果的影响,根据实验结果综合精度和时间损耗指标可以得出,TagRec-UPMF 算法相比当前流行的算法具有一定的优势.

参考文献

[1] RENDLE S, SCHMIDT-THIEME L. Pairwise interaction tensor actorization for personalized tag recommendation [C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2010:81－90.

[2] JÄSCHKE R, MARINHO L, HOTHO A, *et al*. TagRecommendations in folksonomies[J]. Knowledge Discovery in Databases: PKDD,2007,47(2): 506－514.

[3] SIGURBJÖRNSSON B, VAN ZWOL R. Flickr tag recommendation based on collective knowledge[C] //Proceedings of the 17th International Conference on World Wide Web. Beijing: ACM, 2008: 327－336.

[4] SEN S, LAM S K, RASHID A M, *et al*. Tagging, communities, vocabulary, evolution [C]//Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work. New York, USA: ACM, 2006: 181－190.

[5] HOTHO A, JÄSCHKE R, SCHMITZ C, *et al*. BibSonomy: A social bookmark and publication sharing system[C]//Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures. Aalborg, Denmark, 2006:87－102.

[6] **廖志芳,李玲,刘丽敏,等.三部图张量分解标签推荐算法[J].计算机学报**,2012,35(12):2625－2632.
LIAO Zhi-fang, LI Lin, LIU Li-min, *et al*. A tripartite decomposition of tensor for social tagging [J]. Chinese Journal of Computers, 2012,35(12):2625－2632. (In Chinese)

[7] MA H, YANG H, LYU M R, *et al*. Sorec: Social recommend dation using probabilistic matrix factorization [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York, USA: ACM, 2008: 931－940.

[8] SYMEONIDIS P, NANOPOULOS A, MANOLOPOULOS Y. TagRecommendations based on tensor dimensionality reduction[C]//Proceedings of the 2008 ACM Conference on Recommender Systems. Lausanne, Switzerland: ACM, 2008: 43－50.

[9] LANGSETH H, NIELSEN T D. A latent model for collaborative filtering[J]. International Journal of Approximate Reasoning, 2012, 53(4): 447－466.

[10] POLAT H, DU W. SVD-based collaborative filtering with privacy[C]//Proceedings of the 2005 ACM Symposium on Applied Computing. New York, USA: ACM, 2005: 791－795.

[11] MA H, KING I, LYU M R. Learning to recommend with social trust ensemble[C]//Proceedings of the 32nd Inter National ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, USA: ACM, 2009: 203－210.

[12] **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.
ZHU Yu-xiao, LV Lin-yuan. Evaluation metrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163－175. (In Chinese)