科研社交网络中基于联合概率矩阵分解的科技论文推荐方法研究

吴燎原[1] 蒋 军[2] 王 刚[2]

(合肥工业大学科学技术研究院 合肥 230009)[1] (合肥工业大学管理学院 合肥 230009)[2]

摘 要 近年来随着科研社交网络中科技论文数量爆炸式的增长,科研人员很难高效地找到与之相关的科技论文,因此面向科研工作者的科技论文推荐方法应运而生。然而,传统的科技论文推荐方法没有充分挖掘科研社交网络中广泛存在的社会化信息,导致科技论文推荐质量不高。为此,提出了一种科研社交网络中基于联合概率矩阵分解的科技论文推荐方法,在传统概率矩阵分解的基础上,融入了社会化标签信息和社会化群组信息来进行科技论文推荐。为了验证所提方法的有效性,抓取了科研社交网络 CiteULike 上的数据进行了实验。实验结果表明,与其它传统推荐方法相比较,所提方法在 Precision 和 Recall 两个评价指标上均取得了较好的推荐结果,并且能够应用于大规模数据集,具有良好的可扩展性。

关键词 科技论文推荐,科研社交网络,联合概率矩阵分解,推荐方法

中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.042

Study of Scientific Paper Recommendation Method Based on Unified Probabilistic Matrix Factorization in Scientific Social Networks

WU Liao-yuan[1] JIANG Jun[2] WANG Gang[2]

(Institute of Science and Technology, Hefei University of Technology, Hefei 230009, China)[1]

(School of Management, Hefei University of Technology, Hefei 230009, China)[2]

Abstract In recent years, the number of scientific papers in scientific social networks has grown at an explosive rate. It is difficult for researchers to find scientific papers related to their research. Therefore, the paper recommendation for researchers was proposed to solve this problem. However, many problems exist in traditional paper recommendation methods, especially for the fact that a lot of social information in scientific social network are not fully used, resulting in poor quality of paper recommendation. Therefore, this research proposed a new paper recommendation method for researchers in scientific social networks based on the unified probability matrix factorization. This method incorporates social tag information and group information into traditional matrix factorization. In order to verify the validity of the proposed method, we crawled data from a famous scientific social network, i.e. CiteULike, to conduct experiments. Experimental results show that the proposed method gets the best recommendation results at the two evaluation metrics, i.e. Precision and Recall, compared to other traditional recommendation methods. The proposed method is linear with respect to the number of observed data, and performs well in scalability.

Keywords Scientific paper recommendation, Scientific social network, Unified probabilistic matrix factorization, Recommendation method

1 引言

随着互联网技术的普及和飞速发展,互联网已经从 Web1.0 时代步入 Web2.0 时代。Web2.0 时代的最大特点就是互联网上产生了大量由用户自主生成的在线内容,这使得用户不仅是互联网内容的接受者,而且成为了制造者[1]。其中,社交网络作为 Web2.0 时代的一个典型应用,已成为近些年来发展最迅猛的互联网产品之一,使得各种各样的线下关系都可以在网络上建立起来,并在网络上进行交互。随着社交网络的迅速普及,在科研领域也出现了具有其自身特色的社交网络,即科研社交网络,例如 CiteULike、科研之友(ScholarMate)等。科研社交网络的兴起为科研人员提供了一个公共平台,以帮助他们方便地查找和分享感兴趣的科技论文并推广自己的科研成果[2,3]。但是,近些年来科研社交网络中科技论文的数量呈现出爆炸式增长趋势,这使得科研人员很难从海量的科技论文中搜寻到有用的科技论文[3]。因

到稿日期:2016-01-29 返修日期:2016-06-02 本文受国家自然科学基金(71101042,71471054),安徽省自然科学基金(1608085MG150)资助。

吴燎原(1973—),男,硕士,主要研究方向为推荐系统、数据挖掘,E-mail:wliaoyuan@163.com;蒋 军(1991—),男,硕士生,主要研究方向为推荐系统、社交网络分析,E-mail:630740399@qq.com;王 刚(1980—),男,副研究员,主要研究方向为商务智能与商务分析,E-mail:wgedison@gmail.com。

此，如何向科研社交网络中的科研人员推荐符合其特征的科技论文成为一个热点问题[2,3]。

目前，研究者已开始关注面向科研人员的科技论文推荐，并开展了初步的研究工作，现有科技论文推荐方法大致可以分为3类：基于内容的推荐方法[2,3]、协同过滤推荐方法[5-7]和混合推荐方法[8,9]。这些方法虽然在一定程度上可以满足科技论文推荐的要求，但依然存在很多不足之处。首先，在基于内容的推荐方法中，研究者通常只是利用了科技论文的标题、摘要和关键字等容易获取的科技论文内容信息来实现推荐方法[9]。例如，Kim等通过分析关键词语在论文中出现的位置和频率，来推断用户的关键词语偏好和使用模式，从而对每个科研人员做出推荐[4]。其次，在基于协同过滤的科技论文推荐方法中，研究者一般也只是利用了科研人员对科技论文的评分信息，从而得到科研人员对科技论文的预测评分，最终得到面向科研人员的推荐列表。例如，Boger和Bosch将传统的协同过滤算法运用到论文推荐中并且在CiteULike数据集上验证，发现基于用户的协同过滤算法比基于项目的协同过滤算法取得了更好的效果[5]。最后，在混合的科技论文推荐方法中，研究者通常将基于内容的推荐方法和基于协同过滤的推荐方法结合起来[11]。例如，Wang Chong和Blei提出了CTM模型。该模型将传统的协同过滤方法和LDA方法相结合，并根据评价矩阵的稀疏度来改变两者在推荐结果中的比例[8]。虽然混合推荐方法的推荐效果一般优于仅仅使用了单一策略的推荐方法，但是其依然存在数据稀疏性和推荐精度不高等问题。为此，如何利用科研社交网络中的海量社会化信息提升科技论文推荐质量已成为一个重要研究问题。

近些年来，随着科研社交网络的不断发展，在其中聚集了大量科研人员自主生成的社会化信息。在这些社会化信息中有两种信息具有很强的代表性，一种是社会化标签信息。其是由科研人员为自己收藏过的文章所定义的一个或多个描述，科研人员可以通过标签为科技论文分类，也可以通过标签检索具有相同标签的科技论文[2]。所以这些社会化标签从某种程度上可以间接地反映科研人员的兴趣和科技论文的内容。另一种是群组信息。在科研社交网络中，随着科研人员之间交流的不断深入，科研人员自主地在科研社交网络中组建了大量因兴趣爱好相同、研究领域相近的群组，以方便群组成员之间的交流。因此，在相同群组中的科研人员在科研领域和偏好方面具有一定的相似性。虽然上述社会化信息在科研社交网络中普遍存在，但是到目前为止，在科技论文推荐问题中这些社会化信息还没有得到研究者的充分关注，更没有研究者同时将这两种社会化信息引入到科技论文的推荐方法之中。

为了进一步提高科技论文推荐方法的性能，针对以上存在的一些问题，本文提出了一种新的融合社会化标签信息和群组信息的科技论文推荐方法。首先，利用科技论文获得的标签信息和科研人员加入群组的信息，构造科技论文-标签矩阵和基于群组信息的科研人员相关性矩阵。其次，将计算获得的科技论文-标签信息和基于群组信息的科研人员相关性融入到概率矩阵分解方法中，实施联合概率矩阵分解，进而得到潜在科研人员和科技论文特征矩阵。最后，依据潜在科研人员和科技论文特征矩阵计算科研人员对科技论文的预测评分，从而得到面向科研人员的推荐列表。为了验证本文提出的科研社交网络中基于联合概率矩阵分解的科技论文推荐算法的有效性，抓取了科研社交网络CiteULike上的数据进行了实验。实验结果表明，与其它未融入标签信息和基于群组信息的科研人员相关性的同类型推荐方法相比，本文所提出的方法能够在准确率(Precision)和召回率(Recall)两种评价指标上取得更好的结果，从而有效提升了科研社交网络中科技论文推荐的精度。

2 问题形式化定义

本文所提出的推荐算法主要是基于科研人员-科技论文评分矩阵进行相应的计算，从而能够对科研社交网络中科研人员对科技论文的评分进行预测。具体而言，假设科研社交网络中包含 N 个科研人员，其构成科研人员集合 $\mathcal{U}=\{u_1,u_2,\cdots,u_i,\cdots,u_N\}$；包含 M 篇科技论文，其构成科技论文集合 $\mathcal{V}=\{v_1,v_2,\cdots,v_j,\cdots,v_M\}$，以及科研人员对科技论文的评分矩阵 $R=\{R_{i,j}\}_{N\times M}$，其中 $R_{i,j}$ 表示科研人员 u_i 对科技论文 v_j 的评分(比如1到5分)。

目前，基于概率矩阵分解的协同过滤方法是一种应用广泛的评分预测方法[13]。其主要目标是通过概率矩阵分解模型得到科研人员特征矩阵 $U\in R^{K\times N}$ 和科技论文特征矩阵 $V\in R^{K\times M}$，从而使 $U^{\mathrm{T}}V$ 逼近评分矩阵 R。其中，U_i 表示科研人员 u_i 的 K 维特征向量，V_j 表示科技论文 v_j 的 K 维特征向量。根据以上定义，已有评分数据的条件概率有如下定义[13]：

$$p(R|U,V,\sigma_R^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}[N(R_{i,j}|g(U_i^{\mathrm{T}}V_j),\sigma_R^2)]^{I_{i,j}^R} \tag{1}$$

其中，$N(x|\mu,\sigma^2)$ 表示均值为 μ、方差为 σ^2 的高斯分布；$I_{i,j}^R$ 是指示函数，如果科研人员 u_i 对科技论文 v_j 有过评分，则 $I_{i,j}^R=1$，否则 $I_{i,j}^R=0$；$g(x)=1/(1+\exp(-x))$，其目的是将 $U_i^{\mathrm{T}}V_j$ 的值映射到[0,1]区间内。

另外，为了防止过拟合的发生，假设科研人员和科技论文的特征向量服从均值为0的高斯先验，其定义如下：

$$p(U|\sigma_U^2)=\prod_{i=1}^{N}N(U_i|0,\sigma_U^2I) \tag{2}$$

$$p(V|\sigma_V^2)=\prod_{j=1}^{M}N(V_j|0,\sigma_V^2I) \tag{3}$$

根据以上定义，经过贝叶斯推断，可得特征矩阵 U 和 V 的后验概率：

$$\begin{aligned}p(U,V|R,\sigma_R^2,\sigma_U^2,\sigma_V^2)&\propto p(R|U,V,\sigma_R^2)p(U|\sigma_U^2)p(V|\sigma_V^2)\\&=\prod_{i=1}^{N}\prod_{j=1}^{M}[N(R_{i,j}|g(U_i^{\mathrm{T}}V_j),\sigma_R^2)]^{I_{i,j}^R}\times\prod_{i=1}^{N}N(U_i|0,\\&\quad\sigma_U^2I)\times\prod_{j=1}^{M}N(V_j|0,\sigma_V^2I)\end{aligned} \tag{4}$$

综上，概率矩阵分解方法的图模型可以用图1来表示[13]。

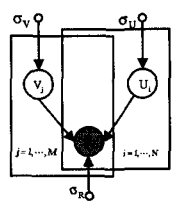

图1 概率矩阵分解图模型

3 基于联合概率矩阵分解的科技论文推荐方法

在传统概率矩阵分解的基础上，本文提出了一种科研社交网络中基于联合概率矩阵分解的科技论文推荐方法，该方法主要包括以下3个部分。

(1)利用科研社交网络中科研人员给科技论文标注的标签信息构造科技论文-标签信息矩阵。同时，利用科研人员加入群组的信息，构造基于群组信息的科研人员相关性矩阵。

(2)将科技论文-标签信息矩阵和科研人员相关性矩阵融入到基于评分信息的概率矩阵分解方法中，实施联合概率矩阵分解方法，得到科研人员特征矩阵和科技论文特征矩阵。

(3)利用已经得到的科研人员特征矩阵和科技论文特征矩阵得到科研人员对科技论文的预测评分，并依据预测评分对科技论文进行排序，将排名靠前的科技论文推荐给科研人员。

为此，本部分首先介绍如何构造科技论文-标签信息矩阵和基于群组信息的科研人员相关性矩阵。其次，详细介绍如何求解科研人员特征矩阵和科技论文特征矩阵的联合概率矩阵分解方法。最后，对所提方法的时间复杂度进行了分析。

3.1 构造科技论文-标签信息矩阵

假设科研社交网络中所有科技论文获得的 W 个标签构成标签集合 $\mathcal{B}=\{b_1,b_2,\cdots,b_k,\cdots,b_W\}$，则可以据此构造科技论文-标签信息矩阵 $F=\{F_{j,k}\}_{M\times W}$。其中，$F_{j,k}$ 表示科技论文 v_j 与标签 b_k 的相关程度。显然科技论文被某一标签标注得越多，则两者相关性越强。具体而言，$F_{j,k}$ 可由式(5)计算得到：

$$F_{j,k}=g(f(v_j,b_k)) \tag{5}$$

其中，$g(x)=1/(1+\exp(-x))$，用于归一化[14]；$f(v_j,b_k)$ 表示科技论文 v_j 被标签 b_k 标注的次数。

3.2 构造基于群组信息的科研人员相关性矩阵

假设科研社交网络中存在 L 个群组构成群组集合 $\mathcal{G}=\{g_1,g_2,\cdots,g_l,\cdots,g_L\}$，则科研人员加入群组的信息可由科研人员-群组相关性矩阵 $A=\{A_{i,l}\}_{N\times L}$ 表示，如果科研人员 u_i 加入了群组 g_l，则 $A_{i,l}=1$，否则 $A_{i,l}=0$。为此，就可以利用科研人员-群组相关性矩阵 A 计算得到科研人员之间的相关性矩阵 $C=\{C_{i,m}\}_{N\times N}$。$C_{i,m}$ 表示科研人员 u_i 和科研人员 u_m 之间的相关程度。显然，两个科研人员共同加入的群组个数越多，他们之间的相关性越强。因此，$C_{i,m}$ 可由式(6)计算得到：

$$C_{i,m}=g(|A_{i\cdot}\cap A_{m\cdot}|) \tag{6}$$

其中，$A_{i\cdot}$ 和 $A_{m\cdot}$ 分别表示科研人员-群组相关性矩阵 A 中的第 i 行和第 m 行向量；$|A_{i\cdot}\cap A_{m\cdot}|$ 表示科研人员 u_i 和科研人员 u_m 共同加入的群组的个数；$g(x)=1/(1+\exp(-x))$，用于归一化。

3.3 基于联合概率矩阵分解的科技论文推荐模型

运用传统的概率矩阵分解方法进行科技论文推荐时，仅仅考虑了科研人员对科技论文的评分信息，却没有考虑科研社交网络上广泛存在的标签信息和科研人员之间的交互信息，很容易造成推荐精度和用户满意度低的问题。为此，本文提出了一种融入社会化标签信息和群组信息的联合概率矩阵分解算法，来向科研社交网络中的科研人员推荐科技论文，其概率图模型如图2所示。其中 B_k 和 Z_m 分别表示标签特征矩阵 B 和潜在因素矩阵 Z 中的特征向量。

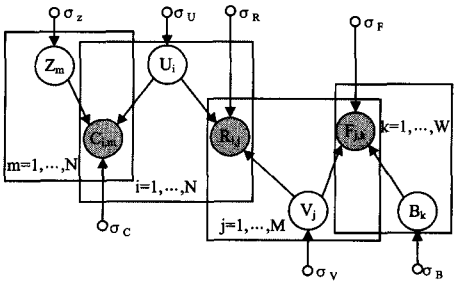

图2 联合概率矩阵分解图模型

在传统概率矩阵分解的基础上，同样基于以下假设：

(1)假设 B_k 和 Z_m 服从均值为0的高斯分布且相互独立，即：

$$p(B|\sigma_B^2)=\sum_{k=1}^{W}N(B_k|0,\sigma_B^2) \tag{7}$$

$$p(Z|\sigma_Z^2)=\sum_{m=1}^{N}N(Z_m|0,\sigma_Z^2) \tag{8}$$

(2)假设科技论文-标签信息矩阵 F 中的元素 $F_{j,k}$ 满足均值为 $g(V_j^TB_k)$、方差为 σ_F^2 的高斯分布且相互独立。因此，科技论文-标签信息矩阵 F 的条件概率分布如下：

$$p(F|V,B,\sigma_F^2)=\sum_{j=1}^{M}\sum_{k=1}^{W}[N(F_{j,k}|g(V_j^TB_k),\sigma_F^2)]^{I_{jk}^F} \tag{9}$$

其中，I_{jk}^F 是指示函数，当科技论文 v_j 被标签 b_k 标注过，则 $I_{jk}^F=1$；否则，$I_{jk}^F=0$。

(3)同理，假设基于群组信息的科研人员相关性矩阵 C 中的元素 $C_{i,m}$ 满足均值为 $g(U_i^TZ_m)$、方差为 σ_C^2 的高斯分布且相互独立。因此，基于群组信息的科研人员相关性矩阵 C 的条件概率分布如下：

$$p(C|U,Z,\sigma_C^2)=\sum_{i=1}^{N}\sum_{m=1}^{N}[N(C_{i,m}|g(U_i^TZ_m),\sigma_C^2)]^{I_{im}^C} \tag{10}$$

其中，I_{im}^C 为指示函数，当科研人员 u_i 和科研人员 u_m 的相关程度为0时，$I_{im}^C=0$；否则，$I_{im}^C=1$。

由图2可知，经过贝叶斯推断可以得到 U,V,B,Z 的后验概率分布，后验分布函数的log函数如下：

$$\begin{aligned}&\ln p(U,V,B,Z|R,F,C,\sigma_U^2,\sigma_V^2,\sigma_B^2,\sigma_Z^2,\sigma_R^2,\sigma_F^2,\sigma_C^2)\\&=-\frac{1}{2\sigma_R^2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}^R(R_{i,j}-g(U_i^TV_j))^2-\frac{1}{2\sigma_F^2}\sum_{j=1}^{M}\sum_{k=1}^{W}I_{jk}^F\\&(F_{j,k}-g(V_j^TB_k))^2-\frac{1}{2\sigma_C^2}\sum_{i=1}^{N}\sum_{m=1}^{N}I_{im}^C(C_{i,m}-g(U_i^TZ_m))^2-\\&\frac{1}{2\sigma_U^2}\sum_{i=1}^{N}U_i^TU_i-\frac{1}{2\sigma_V^2}\sum_{j=1}^{M}V_j^TV_j-\frac{1}{2\sigma_B^2}\sum_{k=1}^{W}B_k^TB_k-\frac{1}{2\sigma_Z^2}\sum_{m=1}^{N}Z_m^T\\&Z_m-\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}^R\ln\sigma_R^2-\sum_{j=1}^{M}\sum_{k=1}^{W}I_{jk}^F\ln\sigma_F^2-\sum_{i=1}^{N}\sum_{m=1}^{N}I_{im}^C\ln\sigma_C^2-\\&K\sum_{i=1}^{N}\ln\sigma_U^2-K\sum_{j=1}^{M}\ln\sigma_V^2-K\sum_{k=1}^{W}\ln\sigma_B^2-K\sum_{m=1}^{N}\ln\sigma_Z^2+c\end{aligned} \tag{11}$$

其中，K 表示特征向量的维度，c 是常量。而最大化式(11)可视为无约束问题，相当于最小化式(12)：

$$E(U,V,B,Z,R,F,C)$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}^{R}(R_{i,j}-g(U_i^TV_j))^2+\frac{\theta_F}{2}\sum_{j=1}^{M}\sum_{k=1}^{W}I_{jk}^{F}(F_{j,k}-g(V_j^TB_k))^2+\frac{\theta_C}{2}\sum_{i=1}^{N}\sum_{m=1}^{N}I_{im}^{C}(C_{i,m}-g(U_i^TZ_m))^2+\frac{\theta_U}{2}\sum_{i=1}^{N}U_i^TU_i+\frac{\theta_V}{2}\sum_{j=1}^{M}V_j^TV_j+\frac{\theta_B}{2}\sum_{k=1}^{W}B_k^TB_k+\frac{\theta_Z}{2}\sum_{m=1}^{N}Z_m^TZ_m \tag{12}$$

其中，$\theta_F=\frac{\sigma_R^2}{\sigma_F^2}$，$\theta_C=\frac{\sigma_R^2}{\sigma_C^2}$，$\theta_U=\frac{\sigma_R^2}{\sigma_U^2}$，$\theta_V=\frac{\sigma_R^2}{\sigma_V^2}$，$\theta_B=\frac{\sigma_R^2}{\sigma_B^2}$，$\theta_Z=\frac{\sigma_R^2}{\sigma_Z^2}$。式(12)的最小值可由梯度下降法求得，参数 U_i，V_j，B_k 和 Z_m 的梯度下降公式如下：

$$\frac{\partial E}{\partial U_i}=\sum_{j=1}^{M}I_{ij}^{R}(g(U_i^TV_j)-R_{i,j})g'(U_i^TV_j)V_j+\theta_C\sum_{m=1}^{N}I_{im}^{C}(g(U_i^TZ_m)-C_{i,m})g'(U_i^TZ_m)Z_m+\theta_UU_i \tag{13}$$

$$\frac{\partial E}{\partial V_j}=\sum_{i=1}^{N}I_{ij}^{R}(g(U_i^TV_j)-R_{i,j})g'(U_i^TV_j)U_i+\theta_F\sum_{k=1}^{W}I_{jk}^{F}(g(V_j^TB_k)-F_{j,k})g'(V_j^TB_k)B_k+\theta_VV_j \tag{14}$$

$$\frac{\partial E}{\partial B_k}=\theta_F\sum_{j=1}^{M}(g(V_j^TB_k)-F_{j,k})g'(V_j^TB_k)V_j+\theta_BB_k \tag{15}$$

$$\frac{\partial E}{\partial Z_m}=\theta_C\sum_{i=1}^{N}(g(U_i^TZ_m)-C_{i,m})g'(U_i^TZ_m)U_i+\theta_ZZ_m \tag{16}$$

3.4 时间复杂度分析

在使用梯度下降法求解时，方法的计算开销主要取决于目标函数 E 和与其相对应的梯度下降公式[15]。由于矩阵 R，C，F 很稀疏，因此很容易得出式(12)中目标函数的时间复杂度为 $O(\rho_RK+\rho_CK+\rho_FK)$，其中 ρ_R，ρ_C，ρ_F 分别表示矩阵 R，C，F 中非零元素的个数。同理，可以得到式(13)－式(16)的时间复杂度。所以，该方法每一次迭代过程的时间复杂度为 $O(\rho_RK+\rho_CK+\rho_FK)$。由此可以看出，方法的时间复杂度随 R，C，F 矩阵中观测数据数量的增加呈线性增长，意味着本文提出的方法可以应用于大规模数据。

4 实验及分析

4.1 实验数据

为了验证本文提出的科研社交网络中基于联合概率矩阵分解的科技论文推荐方法的有效性，运用网络爬虫技术从科研社交网络 CiteULike 中获取了实验所需的原始数据集。原始数据集中包含了科研人员收藏科技论文的信息、科研人员给科技论文标注标签的信息和科研人员加入群组的相关信息等。为了保证实验的有效性，确定了一定的数据筛选规则来进行数据预处理工作。首先，选取了被收藏大于或者等于 2 次的科技论文，在此基础上，进一步选取了收藏科技论文多于 15 篇的科研人员，最后筛选出包含两个及两个以上科研人员的群组。经过上述数据预处理过程，最终获取了包含 1660 个科研人员、70032 篇科技论文、206289 次用户收藏文章的信息、60642 个标签、469 个群组的实验数据集。

4.2 评价指标

本文选用推荐系统领域常用的准确率(Precision)和召回率(Recall)作为评价指标[16,17]，其中，Precision 表示推荐的科技论文中真正符合科研人员兴趣的科技论文所占的比例，如式(17)所示：

$$Precision=\frac{|X\cap Y|}{|X|} \tag{17}$$

其中，X 表示推荐的结果集，Y 表示测试集，Precision 值越大，表明推荐算法的准确率越高。

Recall 表示推荐的科技论文中符合科研人员兴趣的占测试集中所有科技论文的比例，如式(18)所示：

$$Recall=\frac{|X\cap Y|}{|Y|} \tag{18}$$

其中，X 表示推荐的结果集，Y 表示测试集，Recall 值越大，表明推荐算法的精确度越高。

4.3 比较方法和参数设定

为了验证所提方法的有效性，本文在基于内容的推荐方法中选取了 TFIDF 方法作为对比方法，用标签作为科技论文的内容信息[9]；在协同过滤方法中选取了概率矩阵分解方法(PMF)[13]和用户最近邻方法(UserKNN)[17]作为对比方法；在混合推荐方法中，选取了 CTM 方法作为对比方法[8]。另外，为了验证各种社会化信息的重要性，还选取了融入科技论文-标签信息的概率矩阵分解方法(TPMF)和融入基于群组信息的科研人员相关性的概率矩阵分解方法(SPMF)作为对比方法，与本文提出的同时融入科技论文-标签信息和基于群组信息的科研人员相关性的联合概率矩阵分解方法(TSPMF)进行比较。

在实验过程中，随机选择 80% 的实验数据集作为训练集，20% 作为测试集。为了保证实验结果的可靠性，每次实验进行 10 次，最终结果取 10 次实验的平均值[12]。同时，经过实验反复测试，发现参数设定为 $\theta_U=\theta_V=\theta_B=\theta_Z=0.001$，$\theta_F=0.5$，$\theta_C=1$，特征向量的维度 $K=20$ 时，方法效果最优。以下实验若无特别说明，上述所有参数均设定为最优值。

4.4 实验结果与分析

实验首先比较了各种方法在不同推荐个数下的结果。在实验中分别设定推荐个数 $d=5$ 和 $d=10$，具体结果如表 1 所列。

表 1　7 种推荐方法的效果比较

模型	d=5		d=10	
	Precision	Recall	Precision	Recall
TFIDF	0.0568	0.0501	0.0542	0.0567
UserKNN	0.0554	0.0511	0.0546	0.0586
PMF	0.0687	0.0532	0.0632	0.0726
CTM	0.0756	0.0623	0.0716	0.0869
TPMF	0.0739	0.0606	0.0706	0.0870
SPMF	0.0785	0.0621	0.0721	0.0887
TSPMF	0.0824	0.0703	0.0807	0.0940

由表 1 可知，TSPMF 方法取得了比基于内容的推荐方法、协同过滤推荐方法和混合推荐方法更好的实验结果，这表明本文所提方法的有效性。同时，与 TPMF 和 SPMF 相比，TSPMF 方法在 Precision 和 Recall 两个评价指标上的推荐精度更高，表明了同时考虑科技论文-标签信息和基于群组信息的科研人员相关性的重要性。另外，从表 1 中还可以看出，与 TPMF 方法相比，SPMF 在 Precision 和 Recall 两个评价指标上均取得了更好的推荐结果。这一结果表明，在面向科研社交网络的科技论文推荐方法中，基于群组信息的科研人员相关性在提高推荐精度方面的作用大于科技论文-标签信息。

(1)参数 θ_C 对方法的影响

在 TSPMF 方法中，θ_C 可以衡量科研人员受到基于群组信息的用户相关性影响的程度，θ_C 越大，表明基于群组信息的科研人员相关性对方法的作用越大。

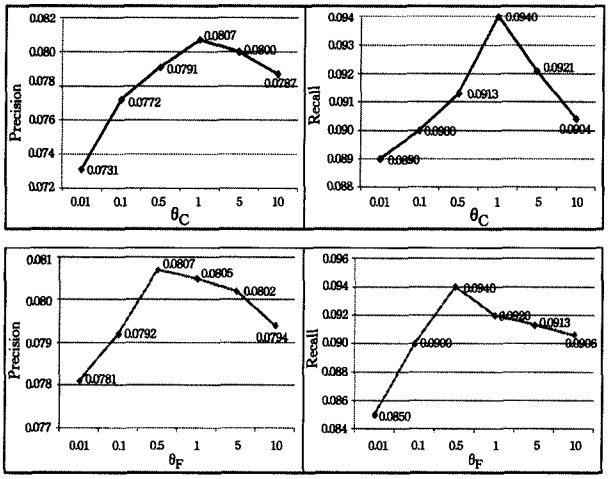

图 3　参数 θ_C 和 θ_F 对实验结果的影响

实验过程中，保持其他参数不变，θ_C 的设定值分别为 0.01，0.1，0.5，1，5，10。此外，设定推荐个数 $d=10$。具体结果如图 3 所示。

图 3 表明参数 θ_C 对 TSPMF 方法有较大的影响。随着 θ_C 的增加，TSPMF 方法在 Precision 和 Recall 两个评价指标上的推荐精度均不断提高，在 $\theta_C=1$ 时取得最大值，这充分说明了通过群组信息获取的科研人员相关性的可靠性，也说明了群组信息的引入对算法的有效性。同时，由图 3 可以发现，当 θ_C 值超过 1 时 TSPMF 方法的推荐效果开始下降，主要是由于 θ_C 过大引起了 TSPMF 方法的过拟合，导致推荐精度降低。

(2)参数 θ_F 对方法的影响

在 TSPMF 方法中，θ_F 可以衡量科技论文受到标签信息影响的程度，其值越大，表明标签信息对方法的作用越大。实验过程中，保持其他参数不变，θ_F 的设定值分别为 0.01，0.1，0.5，1，5，10。此外，设定推荐个数 $d=10$。具体结果如图 3 所示。

图 3 表明参数 θ_F 对 TSPMF 方法的推荐精度有较大的影响。随着 θ_F 的增加，TSPMF 方法在 Precision 和 Recall 两个评价指标上的推荐精度不断提高，在 $\theta_F=0.5$ 时取得最大值，这充分说明了标签信息的有效性。同时，由图 3 还可以发现，当 θ_F 值超过 0.5 时 TSPMF 方法的推荐效果开始下降，主要是由于 θ_F 过大引起了 TSPMF 方法的过拟合，导致推荐精度降低。

结束语　本文对科研社交网络中的科技论文推荐进行了研究。在传统的概率矩阵分解方法的基础上，通过分析用户给科技论文添加的标签信息和用户加入科研群组的信息，提出了一种科研社交网络中基于联合概率矩阵分解的科技论文推荐方法。实验结果表明，与传统的概率矩阵分解方法相比，该方法在多个评价指标下都取得了较好的推荐精度。此外，通过时间复杂度分析表明，本文所提出的方法可应用于大规模数据。今后的工作将更加深入地研究科研社交网络中标签的语义信息和群组的大小对科技论文推荐方法的影响，以期进一步提高该方法的推荐精度和效率。

参 考 文 献

[1] Liu J, Jiang Y, Li Z C, et al. Domain-Sensitive Recommendation with User-Item Subgroup Analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(4): 939-949

[2] Zhao W D, Wu R, Liu H T. Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target[J]. Information Processing and Management, 2016, 26(9): 1-13

[3] Ren Ke, Huang Zhi-xing, Qiu Yu-hui. Interdisciplinary Collaborative Literature Recommendation Based Topic Modeling[J]. Computer Science, 2012, 39(9): 235-239, 261(in Chinese)
任柯，黄智兴，邱玉辉. 基于主题模型的跨学科协作文献推荐[J]. 计算机科学，2012，39(9)：235-239，261

[4] Kim Y S. Text Recommender System Using User's Usage Patterns[J]. Industrial Management & Data Systems, 2010, 111(2): 282-297

[5] Bogers T, Bosch D. Recommending Scientific Articles Using Citeulike[C]// Proc of the ACM Conf on Recommender Systems. New York: ACM, 2008: 287-290

[6] Tian G, Jing L. Recommending Scientific Articles Using Bi-Relational Graph-Based Iterative Rwr[C]// Proc of the 7th ACM Conf on Recommender Systems. 2013: 399-402

[7] Lai C H, Liu D R, Lin C S. Novel personal and group-based trust models in collaborative filtering for document recommendation[J]. Information Sciences, 2013, 239(8): 31-49

[8] Wang C, Blei D M. Collaborative Topic Modeling for Recommending Scientific Articles[C]// Proc of the 17th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining. 2011: 448-456

[9] Sun J S, Ma J, Liu Z Y, et al. Leveraging Content and Connection for Scientific Article Recommendation in Social Computing Contexts[J]. The Computer Journal, 2014, 57(9): 1331-1342

[10] Liang T P, Yang Y F, Chen D N, et al. A Semantic-Expansion Approach to Personalized Knowledge Recommendation[J]. Decision Support Systems, 2008, 45(3): 401-412

[11] Weng S, Chang H. Using Ontology Network Analysis for Research Document Recommendation[J]. Expert Systems with Applications, 2008, 34(3): 1857-1869

[12] Sun Guang-fu, Wu Le, Liu Qi, et al. Recommendation Based on Collaborative Filtering by Exploiting Sequential Behaviors[J]. Journal of Software, 2013, 24(11): 2721-2733(in Chinese)
孙光福，吴乐，刘淇，等. 基于时序行为的协同过滤推荐算法[J]. 软件学报，2013，24(11)：2721-2733

[13] Ma H, Zhou T C, Lyu M R, et al. Improving Recommender Systems by Incorporating Social Contextual Information[J]. ACM Transactions on Information Systems, 2011, 29(2): 1-23

[14] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proc of the ACM Conf. on Recommender Systems. New York: ACM, 2010: 135-142

[15] Tu Dan-dan, Shu Cheng-chun, Yu Hai-yan. Using Unified Probabilistic Matrix Factorization for Contextual Advertisement Recommendation[J]. Journal of Software, 2013, 24(3): 454-464 (in Chinese)
涂丹丹，舒承椿，余海燕. 基于联合概率矩阵分解的上下文广告推荐算法[J]. 软件学报，2013，24(3)：454-464

[16] Jiang M, Cui P, Wang F, et al. Scalable Recommendation with Social Contextual Information [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(11): 2789-2802

[17] 项亮. 推荐系统实践[M]. 北京：人民邮电出版社，2012