计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300121

一种融合科研人员标签的学术论文推荐方法

吴 磊[1] 岳 峰[2] 王含茹[3] 王 刚[3]
1 合肥工业大学人事处 合肥 230009
2 肥工业大学计算机与信息学院 合肥 230009
3 合肥工业大学管理学院 合肥 230009

摘 要 近年来,科研社交网络的兴起在一定程度上转变了科研人员原有的科研交流合作模式,深受科研人员的欢迎;然而,科研社交网络上激增的研究成果数量使得科研人员很难找到自己真正感兴趣的学术论文。因此,为科研人员推荐其感兴趣的学术论文,成为一项重要任务。考虑到科研社交网络中科研人员阅读论文数据的特殊性,文中从单类协同过滤角度考虑科研社交网络中的论文推荐问题。一方面,利用科研人员的标签信息进行更精确的负例抽取,并在此基础上考虑科研人员的活跃度以确定负例数量;另一方面,基于添加完负例的科研人员-学术论文评分矩阵进行概率矩阵分解,在概率矩阵分解阶段融合科研人员标签关联矩阵以及论文相似度信息来进行约束,以缓解数据稀疏对最终结果的不利影响。最后,在科研社交网络"科研之友"上进行实验,采用准确率、召回率、平均准确率、平均倒数排名这4项评价指标对推荐结果的准确性及推荐排序进行验证。实验结果表明,所提方法相较于主流方法取得了更好的结果,在准确率指标上提升了4.19%,验证了所提方法将论文推荐考虑为单类协同过滤问题的有效性,以及社会化信息对推荐的有效辅助作用;并且,所提方法在推荐系统中具有良好的可扩展性,能够在科研社交网络中为科研人员进行有效的论文推荐。

关键词:科研社交网络;论文推荐;单类协同过滤;科研人员标签;概率矩阵分解
中图法分类号 TP391.3

Academic Paper Recommendation Method Combined with Researcher Tag

WU Lei[1], YUE Feng[2], WANG Han-ru[3] and WANG Gang[3]
1 Personnel Department, Hefei University of Technology, Hefei 230009, China
2 School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
3 School of Management, Hefei University of Technology, Hefei 230009, China

Abstract In recent years, the rise of scientific social networks has changed the original mode of exchanges and cooperation among researchers to some extent, which makes scientific social networks well received by researchers. With the surge of research findings on scientific social networks, it's difficult for researchers to find research papers they are really interested in. Consequently, it becomes an important task to recommend the papers that researchers are interested in. Considering the particularity of researchers' reading data, this paper conducted paper recommen-dation from the perspective of one class collaborative filtering. On the one hand, researchers' tag information is used to extract negative cases precisely; on the other hand, based on the researcher-paper matrix with negative instances incorporated, the researchers-tag matrix and papers' similarity information are jointly integrated into the probability matrix factorization, to alleviate the data sparsity problem. Finally, experiments were carried out on a scientific social network, ScholarMate. Four evaluation metrics, namely precision, recall, MAP, and MRR, were adopted to verify the recommendation accuracy as well as the recommendation order. The experimental results show that the proposed method performs better than the baselines with an improvement of 4.19% in terms of the precision, which demonstrate the effectiveness of considering the paper recommendation on scientific social networks as a one-class collaborative filtering problem, the effectiveness of introducing extra social information to improve the recommendation results, and the scalability of the proposed method.

Keywords Scientific social networks, Paper recommendation, One class collaborative filtering, Researcher tag, Probabilistic matrix factorization

到稿日期:2019-03-25 返修日期:2019-09-22 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(71471054,91646111);教育部人文社科基金(18YJC870025);安徽省自然科学基金(1608085MG150)
This work was supported by the National Natural Science Foundation of China (71471054,91646111), Ministry of Education of Humanities and Social Science Foundation (18YJC870025) and Natural Science Foundation of Anhui Province, China (1608085MG150).
通信作者:吴磊(leewu79@hotmail.com)

1 引言

随着信息技术的快速发展，社交网络已开始从服务大众生活逐渐扩展到专业知识分享领域，包括医疗、教育、政治以及学术科研等[1]。其中，科研社交网络作为学术科研领域的一种专业社交网络平台，获得了科研人员的广泛关注[2]。科研人员可以在此分享科研成果、了解其他科研人员的信息、跟踪领域内的最新科研进展等，改变了传统的科研交流合作模式。然而，科研成果数量的激增以及学术质量的混杂，将影响科研人员对科研社交网络的使用，科研人员很难找到自己感兴趣的所有学术论文。因此，为科研人员有效地推荐其感兴趣的学术论文，成为了当前科研社交网络提升服务的关键方向[3]。

已有科研社交网络中的论文推荐方法主要可以分为三大类：基于内容的推荐方法，协同过滤的推荐方法，混合推荐方法[4-7]。基于内容的推荐方法通常收集科研人员历史感兴趣的论文，构建论文内容特征及用户偏好特征，为用户匹配与其历史偏好最相似的论文[5]。例如，Vivacqua 等基于主题模型，融入学术论文内容及引用信息进行推荐[8]。Martin 等利用作者、期刊和关键词等半结构化数据评估论文相似性，并进一步利用这种相似性进行论文推荐[9]。Hong 等提出基于用户文件提取关键词并进行关键词推断的个性化论文推荐方法[10]。然而，基于内容的推荐方法过度依赖于用户自身的历史偏好，缺乏推荐多样性，即它不能推荐出用户感兴趣但是与用户历史浏览论文类型不相似的论文。因此，考虑除用户本身以外的其他用户的意见的影响十分必要。协同过滤的推荐方法就是这样一种方法，该方法认为用户很大程度上会喜欢与自己志趣相投的用户喜欢的东西，这在一定程度上缓解了基于内容推荐方法的局限性。例如，Bogers 等利用传统的协同过滤方法推荐论文，并在此基础上发现基于用户的协同过滤方法比基于项目的协同过滤方法的表现更好[11]。Wang 等利用一种可解释的协同过滤结构和概率模型向科研人员推荐已被其他用户浏览过的论文以及无浏览记录的论文[12]。Lee 等在传统协同过滤方法的基础上，考虑融入用户间的信任信息进行推荐[13]。然而，协同过滤的推荐方法仍然存在局限性，如数据稀疏问题及冷启动问题等，有效评分数据太少以及新用户新物品的加入都是需要解决的问题。混合推荐方法则将基于内容的推荐方法与协同过滤推荐方法相结合，通过对两种方法取长补短来进行论文推荐。例如，Kim 等利用协同过滤方法生成初步推荐列表，再利用基于内容的方法从推荐列表中剔除不相关书籍，从而形成最终的推荐列表[14]。

尽管上述研究已对论文推荐问题有了一定的探索，但是科研社交网络中科研人员对学术论文阅读的反馈信息是二元的，即已读和未读，并不像传统推荐系统具有多等级评分数据。而我们可观测到的数据只是少量可以表明用户正向偏好的数据，其余未标记的数据则是由大量用户不喜欢的数据以及用户可能喜欢但是没看见的数据混合形成，导致整体数据不均衡且稀疏。基于此类数据的特殊性，我们考虑从单类协同过滤(One Class Collaborative Filtering，OCCF)角度对科研社交网络中的学术论文推荐问题进行探索[15-17]。OCCF 问题由 Pan 等提出，其将推荐系统中处理隐式数据的方式定义为单类协同过滤问题，隐式数据包括网页点击、收藏、购买等[15]。之后便有学者开始采用该方法解决推荐问题，例如，Hu 等设置一个相对小的阈值，将小于阈值的用户未选择的项目作为用户不喜欢的项目，通过补充信息进行推荐[18]。Jiang 等在研究微博用户的转发行为时，将夹在两条被转发微博之间的其他未被转发的微博作为用户不感兴趣的微博，在此基础上进行推荐[19]。对于 OCCF 问题来说，不仅原始样本数据非常稀疏，而且样本只有正例，缺少负例，这些问题都会导致最终的预测结果不够精确[19-22]。因此，从负例抽取和解决数据高度稀疏性两方面来提升推荐精度是一个重要方向。

基于以上分析，本文在已有研究的基础上考虑 OCCF 问题，提出了一种融合科研人员标签的学术论文推荐方法(Researcher Tag based One Class Collaborative Filtering，RTOCCF)。利用 OCCF 方法为科研社交网络中的科研人员推荐学术论文，一方面，针对缺少负例问题，在负例抽取阶段融入科研人员标签信息，提升了抽取负例的准确性；另一方面，针对数据稀疏问题，引入概率矩阵分解方法，在此基础上融合学术论文间的相似度以及科研人员-标签关联矩阵，实施联合概率矩阵分解，最终为科研人员进行更精确的学术论文推荐。

2 融合科研人员标签的学术论文推荐

为了对科研社交网络中的科研人员进行学术论文推荐，并提高推荐的精度和质量，本文基于 OCCF 融合科研人员标签信息进行推荐。接下来，本节将分别从问题形式化定义、基于科研人员标签信息的负例抽取、融合科研人员标签的改进 OCCF 方法这 3 个方面对所提方法进行详细的介绍。

2.1 问题的形式化定义

假设在科研社交网络中存在 M 个科研人员 $U=\{u_1, u_2, \cdots, u_i, \cdots, u_M\}$，$N$ 篇学术论文 $V=\{v_1, v_2, \cdots, v_j, \cdots, v_N\}$，$K$ 个科研人员标签 $Q=\{q_1, q_2, \cdots q_k, \cdots q_K\}$。建立科研人员-学术论文历史阅读行为矩阵 $\boldsymbol{R}=\{R_{i,j}\}_{M\times N}$，若科研人员 u_i 阅读过学术论文 v_j，则 $R_{i,j}=1$，反之 $R_{i,j}$ 为空；构建科研人员与其标签之间的关联矩阵 $\boldsymbol{L}=\{L_{i,k}\}_{M\times K}$，若表征科研人员 u_i 的标签中含有标签 q_k，则 $L_{i,k}$ 为 1，否则 $L_{i,k}$ 为 0。考虑学术论文之间的相关性，以 $S_{j,f}$ 表示学术论文 v_f 与 v_j 的相似度。针对本文基于科研人员标签进行学术论文推荐的问题，我们从负例抽取和概率矩阵分解两方面对已有方法进行改进。从负例抽取的角度来说，我们仅能够明确科研人员已阅读过的论文为其真正感兴趣的领域，而不能确定科研人员没有阅读过的论文是其真正不感兴趣的还是实际感兴趣但被遗漏掉的文章。因此，如何从科研人员没有阅读过的文章中准确且合理地选择出负样本进行训练是一个关键问题。同时，由于科研人员标签能够侧面表征科研人员的偏好，本文利用它进行负例抽取之后，还将它融入概率矩阵分解过程以缓解数据稀疏问题。

2.2 融合科研人员标签的负例抽取

由于原始样本稀疏且只有正样本，因此从科研人员未标记的数据中识别出负样本与潜在正样本是一项重要工作。目前已有关于 OCCF 中的负例抽取方法大多采用随机抽取的方式选取负例，即对每个用户来说，他未选择的所有项目都有相同的概率被选作负例，而显然不同项目作为负例的可能性是不同的。针对以上问题，本文在已有方法的基础上利用科研人员的标签信息进行负例抽取。

具体来说，就是利用科研人员的标签表征其偏好倾向，然后计算该偏好与论文内容之间的相似度。相似度越低表示科研人员对这篇论文越不感兴趣，因此，该论文被抽取为负例的可能性越高。我们利用余弦相似度来衡量科研人员与论文之间的相似度：

$$sim(u_i,v_j)=\frac{\boldsymbol{T}_i\cdot\boldsymbol{A}_j}{\|\boldsymbol{T}_i\|\times\|\boldsymbol{A}_j\|} \tag{1}$$

其中，$\boldsymbol{T}_i$ 表示科研人员的标签向量，科研人员的标签为系统根据其研究领域和研究兴趣生成的标签信息，将其表示为一组向量来表征科研人员的偏好倾向；$\boldsymbol{A}_j$ 表示学术论文向量，利用 TF-IDF 方法提取学术论文摘要中的关键词语，从而构成一组向量，与 u_i 科研人员的标签向量进行相似性度量，相似度越高说明科研人员对其越有阅读倾向。负例选取的数目根据科研人员的活跃度来确定：$N_i=\alpha\times\sum_{j=1}^{N}R_{ij}$。其中，科研人员活跃度通过科研人员已阅读过的论文正例数量来表征，已阅读的正例越多，其活跃度越高，对其添加的负例相应地就越多。科研人员 u_i 真正阅读的正例越多，说明 u_i 见过的论文数量越多，其他没有被 u_i 阅读的论文更多的是其看见了但是不喜欢，而不是没看见，因此从 u_i 未阅读的论文中抽取的负例就应该越多。α 是负正例的比例，$\sum_{j=1}^{N}R_{ij}$ 是科研人员 u_i 已阅读的正例数量。另外，添加到矩阵 $\boldsymbol{R}=\{R_{i,j}\}_{M\times N}$ 中的值 $0<r_{i,j}<1$，与科研人员已阅读过的论文相似度越小，就认为该负例是科研人员真正不喜欢的可能性越大，其 $r_{i,j}$ 值也就越接近 0。

基于以上分析，算法 1 给出了基于科研人员标签信息的负例抽取过程。

算法 1　融合科研人员标签信息的负例抽取算法

输入：科研人员-学术论文矩阵 **R**，科研人员标签向量 $\mathbf{T}_i$，学术论文向量 $\mathbf{A}_j$，负例抽取比例 α

输出：添加完负例的科研人员-学术论文矩阵 **R**

1. For i=1,2,…,M
2. 　根据科研人员-学术论文历史阅读行为矩阵 R 中用户 u_i 已选择的正例，确定 u_i 应抽取的负例数 $N_i=\alpha\times\sum_{j=1}^{N}R_{ij}$
3. 　初始化一个负例候选列表 list={0}
4. 　For j=1,2,…,N
5. 　　If(用户 u_i 未选择过 v_j)
6. 　　根据式(1)计算 u_i 与 v_j 的余弦相似度；
7. 　　将其储存为科研人员-学术论文-相似度的键值对 $K(u_i,v_j,sim(u_i,v_j))$；
8. 　　End if
9. 　End for
10. 根据科研人员-学术论文-相似度的键值对 $K(u_i,v_j,sim(u_i,v_j))$ 中的 $sim(u_i,v_j)$，将科研人员 u_i 对应的论文按照相似度从小到大的顺序添加到 list 中，$sim(u_i,v_j)$ 值相同的论文的顺序随机；
11. 从 list 中按顺序选取 N_i 个负例，将 R 中对应位置置换为 $sim(u_i,v_j)$；
12. End for

2.3　融合科研人员标签的 OCCF 推荐方法

为了对已知信息进行有效建模，本文引入概率矩阵分解方法作为基本的推荐框架。概率矩阵分解方法作为近年来推荐领域中较为流行的方法，有着良好的表现。本文利用概率矩阵分解方法，获得两个分别表示科研人员和学术论文的低维潜在特征矩阵，这些特征是刻画科研人员和学术论文的关键因素，最终被用来预测科研人员对学术论文的偏好程度。本文在原有框架的基础上，融入了科研人员标签信息。具体来说，在已添加负例的基础上，一方面，利用科研人员的标签矩阵对分解所得的科研人员潜在特征向量进行约束；另一方面，考虑获取的学术论文间关系，融合文本与时间信息，选择对其有潜在影响的近邻。

针对科研人员潜在特征的获取，主要根据科研人员的标签信息构建科研人员-标签关联矩阵，并使科研人员-标签关联矩阵与科研人员历史阅读行为矩阵同时分解。由于科研人员标签是对科研人员的可视化描述，因此科研人员的潜在特征向量在概率分解过程中显然要受到其标签的影响。

针对学术论文间关系的获取，根据学术论文的文本内容，利用余弦相似度方法计算论文与论文之间的文本相似度：

$$sim(v_j,v_f)=\frac{\boldsymbol{A}_j\cdot\boldsymbol{A}_f}{\|\boldsymbol{A}_j\|\times\|\boldsymbol{A}_f\|} \tag{2}$$

传统的协同过滤方法均忽略了物品被用户需要的时间信息，然而科研人员阅读论文的时间信息可能会隐藏一部分规律，利用这些规律可以在一定程度上挖掘学术论文之间的关系。例如，论文 J 和论文 F 在较短时间内被同一个科研人员阅读，如果这种情况多次出现，那么 J 与 F 很可能存在潜在的影响关系。为了发现这种潜在的关系，将论文近期共同阅读倾向作为权重融入论文相似度计算中。共同阅读倾向表示为：

$$w_{jf}=\frac{T_{jf}}{D(v_j,v_f)} \tag{3}$$

其中，T_{jf} 表示在一定时间内阅读学术论文 v_j 并且阅读过学术论文 v_f 的科研人员的数量，$D(v_j,v_f)$ 表示在一定时间内阅读过学术论文 v_j 的科研人员数量与阅读过学术论文 v_f 的科研人员数量之和。根据共同阅读倾向，将原相似度 $sim(v_j,v_f)$ 优化为 S_{jf}：

$$S_{jf}=\sqrt{w_{jf}\cdot sim(v_j,v_f)} \tag{4}$$

显然，两篇学术论文之间的相似度越高，这两篇学术论文的潜在特征向量应该越相似。在概率矩阵分解阶段，将该相似度作为学术论文间的关联关系融入概率矩阵分解过程，对学术论文潜在特征向量的分解产生一定约束。

综上所述，本文方法的概率图模型如图 1 所示，其中 $\boldsymbol{U}_i$ 表示科研人员在潜在特征空间的分布向量，$\boldsymbol{V}_j$ 表示学术论文在潜在特征空间的分布向量，$\boldsymbol{Q}_k$ 表示标签在潜在特征空间的分布向量，$S_{f,j}$ 表示学术论文 v_f 与 v_j 的综合相似度。

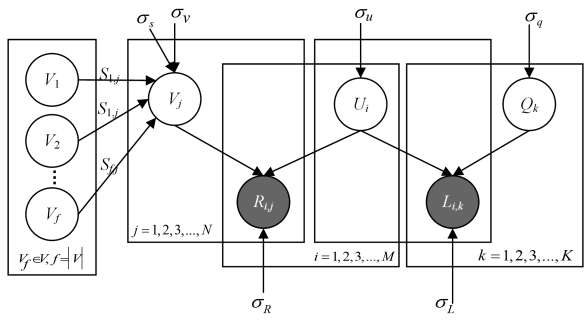

图 1　RTOCCF 概率图模型

Fig. 1　Probabilistic model of RTOCCF

如图1所示，本文提出的RTOCCF方法主要是在对科研人员历史阅读行为矩阵进行分解得到科研人员和学术论文的低维潜在特征矩阵的同时，有效融入论文综合相似度及科研人员-标签关联矩阵，实施联合概率矩阵分解，得到用户潜在特征矩阵 $\boldsymbol{U}\in\boldsymbol{R}^{D\times M}$、项目潜在特征矩阵 $\boldsymbol{V}\in\boldsymbol{R}^{D\times N}$ 和标签潜在特征矩阵 $\boldsymbol{Q}\in\boldsymbol{R}^{D\times K}$，使 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$ 和 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Q}$ 的值尽可能分别逼近科研人员历史阅读行为矩阵 $\boldsymbol{R}$ 和科研人员-标签关联矩阵 $\boldsymbol{L}$。其中，$\boldsymbol{U}_i$ 表示科研人员 u_i 的 D 维特征向量，$\boldsymbol{V}_j$ 表示学术论文 v_j 的 D 维特征向量，$\boldsymbol{Q}_k$ 表示标签 q_k 的 D 维特征向量。根据以上定义，已有科研人员的历史阅读行为矩阵及科研人员-标签关联矩阵的条件概率如下：

$$P(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\sigma_R^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}[N(R_{i,j}\mid g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j),\sigma_R^2)]^{I_{i,j}^R} \tag{5}$$

$$P(\boldsymbol{L}|\boldsymbol{U},\boldsymbol{Q},\sigma_L^2)=\sum_{i=1}^{M}\sum_{k=1}^{K}[N(\boldsymbol{L}_{i,k}\mid g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k),\sigma_L^2)]^{I_{i,k}^L} \tag{6}$$

其中，$N(x|\mu,\sigma^2)$ 表示均值为 μ、方差为 σ^2 的高斯分布；$I_{i,j}^R$ 和 $I_{i,k}^L$ 是指示函数，如果科研人员 u_i 阅读过学术论文 v_j，则 $I_{i,j}^R=1$，否则 $I_{i,j}^R=0$，如果科研人员 u_i 拥有标签 q_k，则 $I_{i,k}^L=1$，否则 $I_{i,k}^L=0$；$g(x)=1/(1+\exp(-x))$，其目的是将 $\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$ 及 $\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k$ 的值映射到[0,1]区间内。

另外，为了防止过拟合，本文假设 $\boldsymbol{U}_i$，$\boldsymbol{V}_j$ 和 $\boldsymbol{Q}_k$ 均服从均值为0的高斯分布且相互独立，其中学术论文的特征向量不仅要服从高斯分布，而且要受到与其相似的学术论文的特征向量的影响，即：

$$P(\boldsymbol{V}|\boldsymbol{S},\sigma_V^2,\sigma_S^2)=\prod_{i=1}^{N}N(\boldsymbol{V}_j|\boldsymbol{0},\sigma_V^2\boldsymbol{I})\times\prod_{i=1}^{N}N(\boldsymbol{V}_j|\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j,\sigma_S^2\boldsymbol{I}) \tag{7}$$

$$P(\boldsymbol{U}|\sigma_U^2)=\prod_{i=1}^{M}N(\boldsymbol{U}_i|\boldsymbol{0},\sigma_U^2\boldsymbol{I}) \tag{8}$$

$$P(\boldsymbol{Q}|\sigma_q^2)=\prod_{k=1}^{K}N(\boldsymbol{Q}_k|\boldsymbol{0},\sigma_q^2\boldsymbol{I}) \tag{9}$$

经过贝叶斯推断，可以得到 $\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}$ 的后验概率分布：

$$\begin{aligned}&P(\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q}|\boldsymbol{R},\boldsymbol{L},\boldsymbol{S},\sigma_R^2,\sigma_u^2,\sigma_v^2,\sigma_L^2,\sigma_q^2,\sigma_s^2)\\&\propto P(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\sigma_R^2)P(\boldsymbol{V}|\boldsymbol{S},\sigma_v^2,\sigma_s^2)P(\boldsymbol{L}|\boldsymbol{U},\boldsymbol{B},\sigma_L^2)P(\boldsymbol{U}|\sigma_u^2)P(\boldsymbol{Q}|\sigma_q^2)\\&=\prod_{i=1}^{M}\prod_{j=1}^{N}[N(\boldsymbol{R}_{i,j}\mid g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j),\sigma_R^2)]^{I_{i,j}^R}\times\prod_{j=1}^{N}N(\boldsymbol{V}_j|\boldsymbol{0},\sigma_v^2\boldsymbol{I})\times\prod_{j=1}^{N}N(\boldsymbol{V}_j|\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j,\sigma_S^2\boldsymbol{I})\times\sum_{i=1}^{M}\sum_{k=1}^{K}[N(\boldsymbol{L}_{i,k}\mid g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k),\sigma_L^2)]^{I_{i,k}^L}\times\prod_{i=1}^{M}N(\boldsymbol{U}_i|\boldsymbol{0},\sigma_u^2\boldsymbol{I})\times\sum_{k=1}^{K}N(\boldsymbol{Q}_k|\boldsymbol{0},\sigma_q^2)\end{aligned} \tag{10}$$

为求解该问题，对式(10)取对数后最小化目标函数：

$$\begin{aligned}&E(\boldsymbol{U},\boldsymbol{V},\boldsymbol{Q},\boldsymbol{R},\boldsymbol{S},\boldsymbol{L})\\&=\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{i,j}^R(R_{i,j}-g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j))^2+\frac{\theta_L}{2}\sum_{i=1}^{M}\sum_{k=1}^{K}I_{i,k}^L(L_{i,k}-g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k))^2+\frac{\theta_S}{2}\sum_{j=1}^{N}(\boldsymbol{V}_j-\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j)^{\mathrm{T}}(\boldsymbol{V}_j-\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j)+\frac{\theta_u}{2}\sum_{i=1}^{M}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i+\frac{\theta_v}{2}\sum_{j=1}^{N}\boldsymbol{V}_j^{\mathrm{T}}\boldsymbol{V}_j+\frac{\theta_q}{2}\sum_{k=1}^{K}\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{Q}_k\end{aligned} \tag{11}$$

其中，$\theta_u=\frac{\sigma_R^2}{\sigma_u^2}$，$\theta_v=\frac{\sigma_R^2}{\sigma_v^2}$，$\theta_L=\frac{\sigma_R^2}{\sigma_L^2}$，$\theta_s=\frac{\sigma_R^2}{\sigma_s^2}$，$\theta_q=\frac{\sigma_R^2}{\sigma_q^2}$，它们反映了各个矩阵对目标函数的影响程度。此优化问题通过梯度下降的方法对目标函数中的变量 $\boldsymbol{U}_i$，$\boldsymbol{V}_j$ 和 $\boldsymbol{Q}_k$ 求偏导，得到目标函数局部最优化的解，进而在多次迭代后得到接近目标函数的全局最优解。参数 $\boldsymbol{U}_i$，$\boldsymbol{V}_j$ 和 $\boldsymbol{Q}_k$ 的梯度计算方法如下：

$$\frac{\partial E}{\partial \boldsymbol{U}_i}=\sum_{i=1}^{M}I_{i,j}^R(g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)-R_{i,j})g'(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)\boldsymbol{V}_j+\theta_L\sum_{k=1}^{K}I_{i,k}^L(g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k)-L_{i,k})g'(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k)\boldsymbol{Q}_k+\theta_v\boldsymbol{V}_j \tag{12}$$

$$\frac{\partial E}{\partial \boldsymbol{V}_j}=\sum_{j=1}^{N}I_{i,j}^R(g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)-R_{i,j})g'(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)\boldsymbol{U}_i+\theta_v\boldsymbol{V}_j+\theta_s(\boldsymbol{V}_j-\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j)-\theta_s\sum_{\langle j|f\in N(j)\rangle}S_{j,f}(\boldsymbol{V}_j-\sum_{f\in N(j)}S_{j,f}\boldsymbol{V}_j) \tag{13}$$

$$\frac{\partial E}{\partial \boldsymbol{Q}_k}=\theta_L\sum_{j=1}^{N}I_{i,k}^L(g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k)-L_{i,k})g'(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{Q}_k)\boldsymbol{U}_i+\theta_q\boldsymbol{Q}_k \tag{14}$$

基于以上分析，算法2给出了本文提出的融合科研人员标签的OCCF推荐方法的详细步骤。

算法2 融合科研人员标签的OCCF推荐方法

输入：矩阵 $\mathbf{R}$ 和 $\mathbf{L}$，论文相似度S，潜在特征维数D，正则化参数 θ_u，θ_v，θ_q，θ_L，θ_s，学习率 λ，最大迭代次数I

输出：科研人员和学术论文的潜在特征矩阵 $\mathbf{U}$ 和 $\mathbf{V}$

1. 调用算法1中已添加过负例的科研人员-学术论文矩阵 $\mathbf{R}$
2. 初始化 $\mathbf{U}$ 和 $\mathbf{V}$，生成随机矩阵 $\mathbf{U}$ 和 $\mathbf{V}$
3. For iter=1,2,…,I do:
4. For each ⟨i,j⟩∈R:
5. 根据式(12)所求梯度更新 $\mathbf{U}_i$：$\mathbf{U}_i=\mathbf{U}_i-\lambda\frac{\partial E}{\partial \mathbf{U}_i}$
6. 根据式(13)所求梯度更新 $\mathbf{V}_j$：$\mathbf{V}_j=\mathbf{V}_j-\lambda\frac{\partial E}{\partial \mathbf{V}_j}$
7. 根据式(14)所求梯度更新 $\mathbf{Q}_k$：$\mathbf{Q}_k=\mathbf{Q}_k-\lambda\frac{\partial E}{\partial \mathbf{Q}_k}$
8. End for
9. End for

3 实验设计

3.1 实验数据

“科研之友”是一个针对学术论文的在线存储、管理和分享平台。在该网站中，科研人员可以阅读和收藏别人上传的论文，并且可以上传自己感兴趣的论文，方便科研人员分享与发现新知识，促进科研进步与交流。本文从“科研之友”网站中抓取数据进行实验，其中包括2532名科研人员及系统生成的标签信息、89781篇学术论文、1017433次科研人员阅读学术论文的信息。

3.2 评价指标

本文采用推荐系统中较为常用的准确率(Precision)、召回率(Recall)、平均准确率(MAP)和平均倒数排名(MRR)4个评价指标，来客观评价本文所提融合科研人员标签的论文推荐方法的有效性。

(1)Precision

准确率是指用户推荐列表中用户真正喜欢的项目占推荐总数的比例[23]。

$$Precision=\frac{|R(u_i)\cap T(u_i)|}{|R(u_i)|} \tag{15}$$

其中，$R(u_i)$ 表示科研人员 u_i 推荐的学术论文集合，$T(u_i)$ 表示测试集中科研人员 u_i 真正阅读过的论文。

(2)Recall

召回率指推荐的学术论文集合中科研人员真正感兴趣的学术论文占测试集中科研人员真正阅读论文数的比例[24]。

$$Recall=\frac{|R(u_i)\cap T(u_i)|}{|T(u_i)|} \tag{16}$$

(3)MAP

在衡量推荐效果时，不仅要衡量算法的准确率，还应该衡量最终推荐给科研人员的论文列表中论文之间的偏序关系，平均准确率就是这样的一个评价指标[23]。

$$MAP=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{|R(u_i)|}\sum_{k=1}^{|R(u_i)|}Pre(R_{ik}) \tag{17}$$

其中，$Pre(R_{ik})$表示科研人员 u_i 的推荐列表 $R(u_i)$中的第 k 篇论文在科研人员 u_i推荐列表和测试集的交集中的位置。

(4)MRR

平均倒数排名是指第一个推荐出的并且符合科研人员偏好的学术论文在推荐列表中的平均位置[25]。

$$MRR=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{rank(F_i)} \tag{18}$$

其中，$rank(F_i)$表示第一个推荐出的并且存在于科研人员 u_i 的测试集中的学术论文在推荐列表中的位置。

3.3　对比方法

由于本文融入科研人员标签对传统 OCCF 法进行改进，因此本次实验选取了以下方法进行对比。

(1)SVD[26]：只使用正例，不考虑添加负例数据，利用 SVD 方法进行建模；

(2)PMF[27]：只使用正例，不考虑添加负例数据，利用 PMF 方法进行建模；

(3)TSPMF：只使用正例，不考虑添加负例数据，并在此基础上利用融入了科研人员标签及论文相似度的 PMF 方法建模；

(4)OCCF-R[28]：考虑随机添加负例数据，并在此基础上利用 PMF 方法建模；

(5)OCCF-N：考虑利用科研人员标签与学术论文相似度添加负例数据，并在此基础上利用 PMF 方法建模；

(6)SOCCF：考虑利用科研人员标签与学术论文相似度添加负例数据，并在此基础上利用融入了论文相似度的 PMF 方法建模；

(7)RTOCCF：考虑利用科研人员标签与学术论文相似度添加负例数据，并在此基础上利用融入了科研人员标签及论文相似度的 PMF 方法建模。

3.4　实验流程

本实验使用 Hold-Out 方法进行验证，随机选取原始数据集的 80%为训练集，剩下的 20%为测试集。实验结果取 10 次实验的平均值，经过反复测试，本文将参数设定为推荐个数 $n=10$、特征向量的维度 $D=40$、添加负例时的负例比例 $\alpha=15$、矩阵分解迭代次数 $I=150$ 时，方法效果最优。在下文的实验中若非特别说明，上述所有参数均设定为最优值。

4　结果分析与讨论

4.1　实验结果

根据上述实验设计，本文对“科研之友”数据集上的实验结果进行统计，取推荐个数 $n=10$，特征维度 $D=40$，得到如表 1 所列的实验结果。其中，每行分别表示当前对比方法在 Precision，Recall，MAP，MRR 这 4 个评价指标下的结果。

表 1　RTOCCF 与对比算法推荐结果的比较

Table 1　Comparison of RTOCCF and other methods

（单位：%）

模型	*Precision*	*Recall*	*MAP*	*MRR*
SVD	6.28	5.47	18.21	17.34
PMF	8.56	7.73	20.26	19.86
TSPMF	10.25	8.51	21.73	20.81
OCCF-R	9.93	8.49	21.90	20.63
OCCF-N	10.14	8.62	21.98	20.76
SOCCF	10.73	9.01	22.80	21.15
RTOCCF	**11.18**	**9.87**	**23.44**	**21.68**

由表 1 可以看出，RTOCCF 方法在 Precison，Recall，MAP，MRR 这 4 个评价指标下均优于其他 6 种对比方法，证明了本文提出的融合科研人员标签的论文推荐方法的有效性。从表中还可以看出，基于 OCCF 的方法整体优于未考虑 OCCF 的方法，即 OCCF-R，OCCF-N，SOCCF，RTOCCF 在 4 个评价指标下的结果在大部分情况下优于 SVD，PMF，SPMF 方法的结果，且相对于 SVD 和 PMF 两种方法均取得了较大幅度的提高。对比 TSPMF 与 RTOCCF 可以发现，RTOCCF 在 4 个评价指标下的结果均优于 TSPMF，证明了考虑单类协同过滤的重要性。对比 RTOCCF，OCCF-R，OCCF-N，SOCCF 方法可以看出，负例抽取与融合信息均对提高推荐结果有显著效果。综合以上分析，对于 OCCF 数据而言，利用科研人员标签与学术论文相似度添加负例数据，以及在概率矩阵分解过程中融入科研人员标签和论文相似度均有助于提高推荐质量，这也进一步验证了本文提出的同时考虑负例抽取和融合科研人员标签的 RTOCCF 方法的有效性。

4.2　实验分析与讨论

由于实验中推荐个数 n、潜在特征维度 D、负例抽取比例 α 和迭代次数 I 都是会对推荐效果产生影响的重要因素，因此接下来本节分别讨论这 4 个参数对实验结果的影响。

(1)推荐个数 n 对实验结果的影响

为了验证实验过程中推荐个数 n 对各方法实验结果的影响，保持其他参数为最优值不变，设置推荐个数 n 分别为 10，20，30，40，50。考虑分别在 Precision 和 MAP 两个评价指标下不同推荐个数对实验结果的影响，具体实验结果如图 2 和图 3 所示。

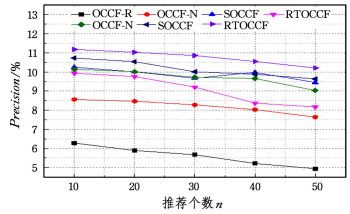

图 2　不同推荐个数 n 下的 *Precision* 比较

Fig. 2　*Precision* under different recommendation number n

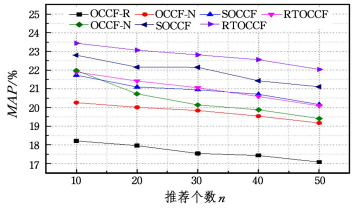

图 3　不同推荐个数 n 下的 *MAP* 比较

Fig. 3　*MAP* under different recommendation number n

由图 2 和图 3 可以发现，与其他方法相比，本文所提

RTOCCF 方法在所有推荐个数下均取得了最好的实验结果，证明了新方法的有效性。另外，从图中可以看出，随着推荐个数的增加，推荐的 *Precision* 和 *MAP* 逐渐减小，即在推荐个数为 10 时结果最优。这可能是由于随着推荐个数的增加，当推荐列表达到一定长度时，科研人员对位于推荐列表尾部的学术论文的偏好程度差距不大，因此最终选取 $n=10$ 作为最优推荐个数。

(2)潜在特征维度 D 对实验结果的影响

潜在特征维度 D 的选取对推荐结果非常重要。如果 D 选择得过小，科研人员和学术论文的隐式特征就不能很好地在隐式空间中有效地表现出来；相反，如果 D 取值过大，那么计算复杂性就会大幅增加，并且会造成学习的过拟合。为了有效验证特征向量维度 D 对各方法实验结果的影响，保持其他参数为最优值不变，设置特征向量维度 D 为 5，10，20，40，具体实验结果如图 4 和图 5 所示。

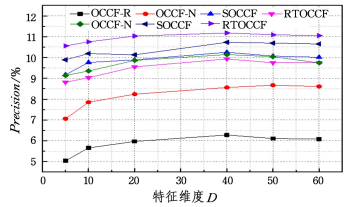

图 4 不同特征维度 D 下的 *Precision* 比较

Fig. 4 *Precision* under different factor dimension D

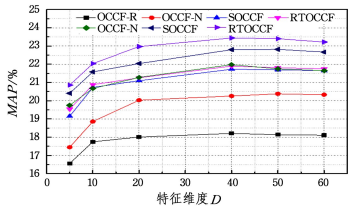

图 5 不同特征维度 D 下的 *MAP* 比较

Fig. 5 *MAP* under different factor dimension D

由图 4 和图 5 可以发现，特征向量维度 D 对各方法有着重要的影响。当潜在特征维度为 5～10 时，*Precision* 和 *MAP* 值上升得很快，而在特征维度为 10～40 时，*Precision* 和 *MAP* 值的上升速度变缓。这可能是由于潜在特征维度为 5 时，数量过少，不足以表征出科研人员和学术论文的很多潜在特征；D 的增大，不仅增加了有效潜在特征，而且也增加了噪声，因此涨幅逐渐变缓。从图中可以看出，当 D 为 40 时，推荐效果达到最优；而当 D 增加到 50 及以后，结果逐渐趋于平稳甚至下降。因此，本文选取 $D=40$ 为最优潜在特征维数。

(3)负例抽取比例 α 对实验结果的影响

在本文所提出的负例抽取方法中，负例比例 α 是重要的影响因素。图 6 和图 7 给出了本文方法及对比方法在不同 α 下的 *Precision* 和 *MAP* 的比较结果。从图 6 和图 7 中可以看出，当 α 较小时，随着 α 的增大，推荐结果的 *Precision* 和 *MAP* 都不断提高，在 α 大约为 15 时最高，但当 α 大于 15 以后 *Precision* 和 *MAP* 开始趋于平缓或降低。这是因为过多的负例同样会造成训练结果趋于负向，影响结果的区分度。另外，随着 α 的增大，需要训练的样本数目也会增多，计算量也会相应增加。从图中还可以看出，无论 α 的取值如何，本文方法都优于其他对比方法，从而证明了本文方法的有效性。

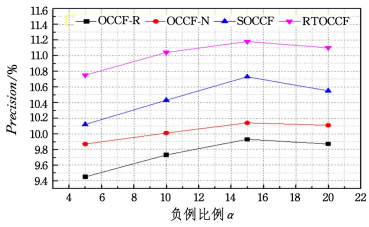

图 6 不同负例比例 α 下的 *Precision* 比较

Fig. 6 *Precision* under different negative proportion α

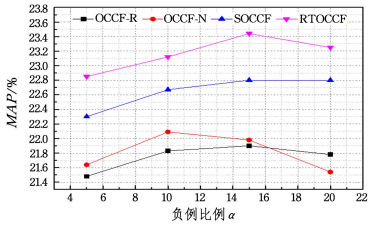

图 7 不同负例比例 α 下的 *MAP* 比较

Fig. 7 *MAP* under different negative proportion α

(4)迭代次数 I 对实验结果的影响

在矩阵分解过程中，迭代次数也是一个重要影响因素，迭代次数过小会导致结果不能收敛，迭代次数设置得过大则会增加计算复杂度。因此，本文对矩阵分解中的迭代次数对推荐结果的影响进行分析与验证，保持其他参数为最优值不变，设置迭代次数 I 为 50，100，150，200，具体实验结果如图 8 和图 9 所示。

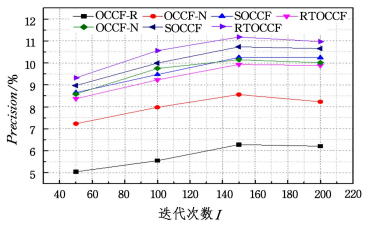

图 8 不同迭代次数 I 下的 *Precision* 比较

Fig. 8 *Precision* under different iteration number I

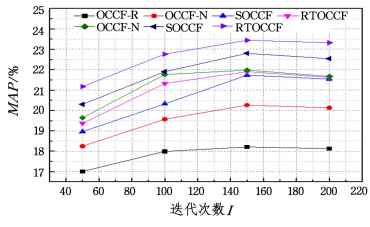

图 9 不同迭代次数 I 下的 *MAP* 比较

Fig. 9 *MAP* under different iteration number I

由图 8 和图 9 可以发现，迭代次数从 50 增加到 100 的过程中，*Precision* 和 *MAP* 的上升幅度较大，这可能是由于迭代次数为 50 时，次数过少，使结果未能达到局部最优解，而随着迭代次数增加到 150 及 200，推荐结果逐渐达到收敛，且在迭代次数 $I=150$ 时结果最优。因此，本文最终选取 $I=150$ 作为最优迭代次数。

结束语 本文针对科研社交网络对科研人员的论文推荐问题，提出了一种融合科研人员标签的论文推荐方法 RTOCCF。首先，根据科研人员标签与未读论文之间的相似度抽取添加负例数据；然后，融合论文相似度及科研人员标签信息进

行联合概率矩阵分解；最后，在科研社交网络“科研之友”上进行了实验，验证了本文方法的有效性。在今后的研究中，我们将进一步考虑科研社交网络中包含的大量其他附加信息，如时间信息等；并且针对 OCCF 问题中的数据不平衡问题，我们可以考虑更有效的方法（如迭代更新）以更好地解决该问题。

参 考 文 献

[1] LI L L, WU X N. A Study of Scientific Social Network's Development and Trend. Research on Library Science[J]. Research on Library Science, 2013, 10(1): 36-41.

[2] WEI J, HE D Q. User participation in an academic social networking service: A survey of open group users on M endeley [J]. Journal of the American Society for Information Science and Technology, 2015, 66(5): 890-904.

[3] SUGIYAMA K, KAN M Y. Scholarly paper recommendation via user's recent research interests[C]// Proceedings of the 10th annual joint conference on Digital libraries. Gold Coast, Queensland, Australia, 2010: 29-38.

[4] PANDEY A K, RAJPOOT D S. Resolving Cold Start problem in recommendation system using demographic approach[C]// 2016 International Conference on Signal Processing and Communication (ICSC). IEEE, 2016: 213-218.

[5] WANG G, HE X, ISHUGA C I J K. HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network[J]. Knowlwdge Based Systems, 2018, 148(9): 85-99.

[6] PHILIP S, SHOLA P B, JOHN A O. Application of content-based approach in research paper recommendation system for a digital library[J]. International Journal of Advanced Computer Science and Applications, 2014, 5(10): 37-40.

[7] TSOLAKIDIS A. Research publication recommendation system based on a hybrid approach[C]// Proceedings of the 20th Pan-Hellenic Conference on Informatics. ACM, New York, USA, 2016: 78.

[8] VIVACQUA A S, OLIVEIRA J, DE SOUZA J M. i-ProSE: inferring user profiles in a scientific context[J]. The Computer Journal, 2009, 52(7): 789-798.

[9] MARTÍN G H, SCHOCKAERT S, CORNELIS C, et al. Using semi-structured data for assessing research paper similarity[J]. Information Sciences, 2013, 221(35): 245-261.

[10] HONG K, JEON H, JEON C. Personalized research paper recommendation system using keyword extraction based on user-profile[J]. Journal of Convergence Information Technology, 2013, 8(16): 106.

[11] BOGERS T, VAN DEN BOSCH A. Recommending scientific articles using citeulike[C]// Proceedings of the 2008 ACM Conference on Recommender Systems. Lausanne, Switzerland, 2008: 287-290.

[12] WANG C, BLEI D M. Collaborative topic modeling for recommending scientific articles[C]// Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, USA, 2011: 448-456.

[13] LEE D H, BRUSILOVSKY P. Using self-defined group activities for improvingrecommendations in collaborative tagging systems[C]// Proceedings of the fourth ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 221-224.

[14] KIM H K, OH H Y, GU J C, et al. Commenders: A recommendation procedure for online book communities[J]. Electronic Commerce Research and Applications, 2011, 10(5): 501-509.

[15] PAN R, ZHOU Y, CAO B, et al. One-class collaborative filtering [C]// 2008 Eighth IEEE International Conference on Data Mining. Antwerp, Belgium, 2008: 502-511.

[16] SUN J, WANG G, CHENG X, et al. Mining affective text to improve social media item recommendation[J]. Information Processing & Management, 2015, 51(4): 444-457.

[17] PAPPAS N, POPESCU-BELIS A. Sentiment analysis of user comments for one-class collaborative filtering over ted talks [C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 773-776.

[18] HU Y, KOREN Y, VOLINSKY C. Collaborative filtering for implicit feedback datasets[C]// 2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008: 263-272.

[19] JIANG M, CUI P, LIU R, et al. Social contextual recommendation[C]// Proceedings of the 21st ACM international conference on Information and knowledge management. Maui, HI, USA, 2012: 45-54.

[20] CHEN K, CHEN T, ZHENG G, et al. Collaborative personalized tweet recommendation[C]// Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Portland, Oregon, USA, 2012: 661-670.

[21] LI Y, HU J, ZHAI C X, et al. Improving one-class collaborative filtering by incorporating rich user information[C]// Proceedings of the 19th ACM international conference on Information and knowledge management. Toronto, Ontario, Canada, 2010: 959-968.

[22] KAYA H, ALPASLAN F N. Using social networks to solve data sparsity problem in one-class collaborative filtering[C]// 2010 Seventh International Conference on Information Technology: New Generations. Las Vegas, NV, USA, 2010: 249-252.

[23] RESNICK P, VARIAN H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-59.

[24] SCHAFER J B, FRANKOWSKI D, HERLOCKER J, et al. Collaborative filtering recommender systems[M]// The adaptive web. Berlin: Springer, 2007: 291-324.

[25] LI H. Learning to rank for information retrieval and natural language processing[J]. Synthesis Lectures on Human Language Technologies, 2011, 4(1): 1-113.

[26] ZHANG S, WANG W, FORD J, et al. Using singular value decomposition approximation for collaborative filtering[C]// Seventh IEEE International Conference on E-Commerce Technology (CEC'05). Munich, Germany, 2005: 257-264.

[27] MNIH A, SALAKHUTDINOV R R. Probabilistic matrix factorization[C]// Advances in neural information processing systems. Vancouver, BC, Canada, 2008: 1257-1264.

[28] PAPPAS N, POPESCU-BELIS A. Adaptive sentiment-aware one-class collaborative filtering[J]. Expert Systems with Applications, 2016, 43(9): 23-41.

WU Lei, born in 1979, Ph.D, doctorial student, lecturer. His main research interests include recommender system and so on.