

# 一种集成链接和属性信息的社区挖掘方法

贺超波<sup>1),(2)</sup> 汤 庸<sup>2)</sup> 刘 海<sup>2)</sup> 赵淦森<sup>2)</sup> 陈启买<sup>2)</sup> 黄昌勤<sup>2)</sup>

<sup>1)</sup>(仲恺农业工程学院信息科学与技术学院 广州 510225)

<sup>2)</sup>(华南师范大学计算机学院 广州 510631)

**摘 要** 现有复杂网络社区挖掘方法由于单一利用节点链接信息或属性信息,从而无法有效发现成员链接紧密且属性高度相同的社区,针对该问题提出一种可集成节点链接和属性信息进行社区挖掘的方法:LANMF. LANMF基于非负矩阵分解模型,以联合矩阵分解的形式统一分解复杂网络节点链接矩阵以及属性关联矩阵,可直接获得节点与社区归属关系矩阵以及属性与社区关联矩阵,社区成员在链接结构紧密度以及属性相关性上可得到很好的保证.设计了乘性迭代更新规则作为联合矩阵分解优化算法并从数学上严格证明了其正确性和收敛性.实验结果表明:LANMF的社区挖掘质量优于现有典型的同类社区挖掘方法,能直接有效挖掘社区,而且实际应用表明LANMF适合用于挖掘现实世界复杂网络中的主题社区以及重叠社区.

**关键词** 社区挖掘;非负矩阵分解;复杂网络;链接信息;属性信息

**中图法分类号** TP391 **DOI号** 10.11897/SP.J.1016.2017.00601

## Method for Community Mining Integrating Link and Attribute Information

HE Chao-Bo<sup>1),(2)</sup> TANG Yong<sup>2)</sup> LIU Hai<sup>2)</sup> ZHAO Gan-Sen<sup>2)</sup>  
CHEN Qi-Mai<sup>2)</sup> HUANG Chang-Qin<sup>2)</sup>

<sup>1)</sup>(School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225)

<sup>2)</sup>(School of Computer Science, South China Normal University, Guangzhou 510631)

**Abstract** Most of existing methods for community mining in complex network only consider the usage of link information or attribute information, and thus are not effective enough to discover high quality community in which members link to each other densely and have nearly the same attributes. Aiming at this problem, we propose a method named LANMF that can integrate together link and attribute information to mine community in complex network. LANMF is based on nonnegative matrix factorization (NMF) model and can factorize uniformly nodes link matrix and attributes association matrix using the form of joint matrix approximating factorization. By using the optimization solution this method can directly obtain the community membership matrix and the associated strength matrix of the attribute and community. Furthermore, relevancies of nodes link structure and attributes in each community can be well guaranteed. LANMF uses multiplicative iterative update rules as the joint matrix factorization optimization algorithm, whose correctness and convergence are strictly proven. Furthermore, extensive experimental results show that the

收稿日期:2015-11-20;在线出版日期:2016-05-08. 本课题得到国家“八六三”高技术研究发展计划项目基金(2013AA01A212)、国家自然科学基金(61370178, 61370229)、广东省自然科学基金(S2012030006242, 2015A030310509)、广东省科技计划项目(2016A030303058, 2016A020210130, 2015A020209178, 2015B010129009, 2014B010103004, 2014B010117007, 2015A030401087, 2015B010110002)、广州市云计算安全与测评技术重点实验室开放基金(GZCSKL-1407)资助. 贺超波,男,1981年生,博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为数据挖掘与社会计算. E-mail: hechaobo@foxmail. com. 汤 庸(通信作者),男,1964年生,博士,教授,博士生导师,主要研究领域为云计算和大数据. E-mail: ytang4@qq. com. 刘 海,男,1974年生,博士,副教授,主要研究方向为机器学习和语义网. 赵淦森,男,1977年生,博士,教授,主要研究领域为云计算与网络安全. 陈启买,男,1965年生,硕士,教授,主要研究领域为数据仓库与数据挖掘. 黄昌勤,男,1972年生,博士,教授,主要研究领域为服务计算和语义网.

quality of community mining using LANMF is better than that of state-of-the-art methods and it can mine community directly and effectively. Moreover, practical application cases show that LANMF is suitable to mine topic community and overlapped community in the real world complex network.

**Keywords** community mining; nonnegative matrix factorization; complex network; link information; attribute information

1 引 言

现实世界中的复杂网络,如社交网络、科技文献合著网络以及蛋白质交互网络等,不仅存在小世界、无标度等特性,而且还具有明显的社区结构,相同社区的节点链接紧密,不同社区的节点链接稀疏.对复杂网络进行社区挖掘对于学习网络拓扑结构特征、功能特征以及预测网络演变趋势等具有重要作用<sup>[1]</sup>.目前,社区挖掘已成为复杂网络科学的研究热点,众多研究人员提出了许多不同类型的社区发现方法,其中包括基于图分割的社区挖掘方法<sup>[2-4]</sup>、基于模块度函数优化的社区挖掘方法<sup>[5-8]</sup>、基于概率模型的社区挖掘方法<sup>[9-11]</sup>、基于聚类算法的社区挖掘方法<sup>[12-13]</sup>、基于生物智能算法的社区挖掘方法<sup>[14-16]</sup>以及基于多 Agent 的自组织社区挖掘方法<sup>[17-18]</sup>等类型.这些方法均能够有效提取复杂网络的社区结构,但由于普遍只考虑利用网络节点的链接信息或属性信息,往往会得到不理想的社区挖掘结果.例如对图 1 所示的社交网络示例进行社区挖掘,该社交网络包含用户好友链接信息以及用户研究兴趣属性信息:大数据和社交网络,在设置社区挖掘数为 2 的情况下,图 2、图 3 分别为只考虑链接信息和属性信息的社区挖掘结果,图 4 则为综合考虑用户链接信息和属性信息的社区挖掘结果.可以看出基于用户链接信息的社区挖掘结果中,社区 1 的成员虽然链接紧密,但属性有很大不同,另外基于用户属性信息

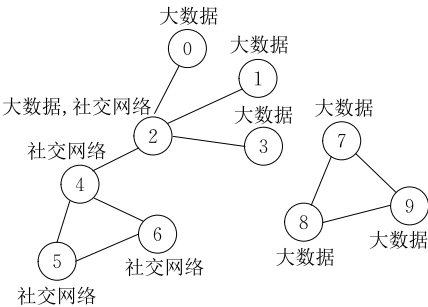


图 1 社交网络示例

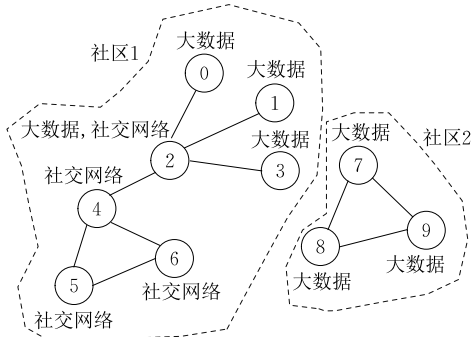


图 2 基于用户链接信息的社区挖掘

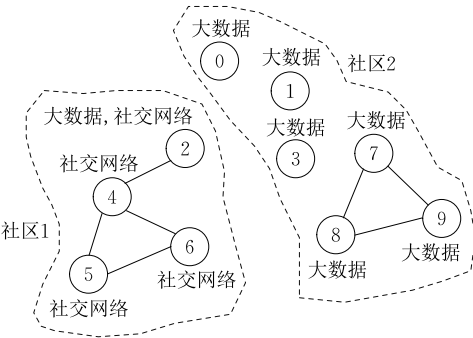


图 3 基于用户属性信息的社区挖掘

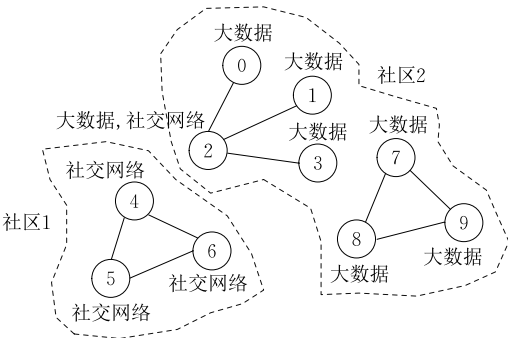


图 4 集成用户链接和属性信息的社区挖掘

的社区挖掘结果中,社区 1 的成员虽然属性相同,但部分成员丢失了链接信息,而集成用户链接和属性信息的社区挖掘结果中,各个社区内的成员不仅链接紧密,而且属性高度相同.理想的社区挖掘结果不仅社区具有紧密的内部链接结构,而且成员节点具有高度相同的属性,这对于在社交网络中进行主题

社区发现以及进行基于社区的社会化推荐等应用具有重要意义,所以有必要研究能够集成节点链接和属性信息进行社区挖掘的方法。

目前,集成节点链接和属性信息进行社区挖掘已有相关研究,具有代表性的工作分别有 Cheng 等人<sup>[19]</sup>提出的 SACcluster 方法, Xu 等人<sup>[20]</sup>提出的 BAGC 方法以及 Ruan 等人<sup>[21]</sup>提出的 CODICIL 方法,但总的来说,以上方法对链接信息和属性信息需要分别设计不同模型处理,社区挖掘过程需要划分多个阶段,挖掘过程复杂化从而不适合实际应用。针对该问题,本文则利用非负矩阵分解(Nonnegative Matrix Factorization, NMF)模型适合直接对复杂网络数据进行聚类以及具有强大的聚类结果可解释能力的特点,提出了一种基于 NMF 的可统一集成链接和属性信息的社区挖掘方法:LANMF,所做的主要工作包括:(1) LANMF 方法基于非负矩阵联合分解模型统一集成节点链接和属性信息,并能够通过矩阵近似分解方法得到节点与社区之间的归属关系矩阵以及各社区与属性之间的关联强度矩阵,可以利用分解矩阵信息直接挖掘社区;(2) 利用乘性迭代更新规则实现对社区挖掘矩阵分解模型的优化求解,从理论上严格证明了相关求解规则的正确性和收敛性,并设计了具有线性时间复杂度的社区挖掘算法;(3) 进行了大量对比实验,结果表明 LANMF 比现有代表性方法具有更好的社区挖掘质量。此外,具体应用实例表明 LANMF 适合对复杂网络进行主题社区以及重叠社区挖掘。

本文第 2 节介绍 NMF 的基本模型、基于 NMF 的社区挖掘以及集成链接和属性信息的社区挖掘研究现状;第 3 节给出本文社区挖掘问题的描述并详细介绍 LANMF 相关问题的解决方法;相关实验及应用实例分析在第 4 节介绍;第 5 节给出本文工作的总结与展望。

## 2 相关工作

### 2.1 NMF 基本模型

由 Lee 和 Seung<sup>[22]</sup>提出的 NMF 是一种低秩矩阵近似逼近模型,适合分析非负值矩阵的主要组成部分,在数学上该模型可以描述为:给定一个矩阵  $\mathbf{X} \in \mathbb{R}_+^{m \times n}$  和期望的秩  $k \leq \min(m, n)$ , 可以将  $\mathbf{X}$  分解为两个矩阵  $\mathbf{P} \in \mathbb{R}_+^{m \times k}$  和  $\mathbf{Q} \in \mathbb{R}_+^{k \times n}$ , 且  $\mathbf{X} \approx \mathbf{PQ}$ , 其中  $\mathbb{R}_+$  为非负值元素集合,  $\mathbf{P}$  称为基矩阵,  $\mathbf{Q}$  为系数矩阵, 矩阵  $\mathbf{X}$  中的每一个列向量  $\mathbf{X}_i$  可以通过  $\mathbf{P}$  以一

定的权值来表示,加权系数则为  $\mathbf{Q}$  中对应列向量  $\mathbf{Q}_i$  所包含的元素,即  $\mathbf{X}_i = \mathbf{PQ}_i$ 。  $\mathbf{P}$  和  $\mathbf{Q}$  可通过求解以下目标函数获得

$$\min_{\mathbf{P} \geq 0, \mathbf{Q} \geq 0} F(\mathbf{X}, \mathbf{PQ}) \quad (1)$$

目标函数  $F(\mathbf{X}, \mathbf{PQ})$  用于量化矩阵  $\mathbf{X}$  和  $\mathbf{PQ}$  的近似逼近误差, Lee 和 Seung 在文献[23]提出两种目前 NMF 广泛使用的目标函数:

(1) 基于欧氏距离的目标函数:

$$F(\mathbf{X}, \mathbf{PQ}) = \|\mathbf{X} - \mathbf{PQ}\|_F^2 \quad (2)$$

其中  $\|\cdot\|_F^2$  表示矩阵的 Frobenius 范数, 当且仅当  $\mathbf{X} = \mathbf{PQ}$  时, 式(2)获得最小值 0。

(2) 基于 Kullback-leibler 离散度的目标函数:

$$F(\mathbf{X}, \mathbf{PQ}) = D(\mathbf{X} \parallel \mathbf{PQ}) \\ = \sum_{ij} \left( \mathbf{X}_{ij} \log \frac{\mathbf{X}_{ij}}{(\mathbf{PQ})_{ij}} - \mathbf{X}_{ij} + (\mathbf{PQ})_{ij} \right) \quad (3)$$

同理, 当且仅当  $\mathbf{X} = \mathbf{PQ}$  时, 式(3)获得最小值 0。文献[23]指出上述两个目标函数对于单独的  $\mathbf{P}$  或  $\mathbf{Q}$  均是凸函数, 但是同时对于  $\mathbf{P}$  和  $\mathbf{Q}$  却不是凸函数, 要找到上述两个函数的全局最优解是 NP 难问题。现实中, 往往使用优化算法得到局部最优解作为矩阵分解结果, 较为常用的方法为迭代更新求解<sup>[24-25]</sup>, 但需要证明迭代更新规则可以保证  $F$  为非增函数, 且可以收敛获得最小值。文献[23]提出为  $F$  设计辅助函数以及进行迭代更新规则收敛性证明的思路, 并给出如下需满足的定义和引理。

**定义 1.**  $G(h, h')$  是  $F(h)$  的辅助函数, 当且仅当满足:  $G(h, h') \geq F(h)$  和  $G(h, h) = F(h)$ 。

**引理 1.** 如果  $G$  是辅助函数, 那么  $F$  在式(4)表示的迭代更新规则上为非增函数, 其中  $h'$  为第  $t$  次迭代更新后的  $h$  值。

$$h^{t+1} = \arg \min_h G(h, h') \quad (4)$$

证明.  $F(h^{t+1}) \leq G(h^{t+1}, h') \leq G(h^t, h') = F(h^t)$ 。

注意到只有当  $h'$  是  $G(h, h')$  的局部极小点时有  $F(h^{t+1}) = F(h^t)$ , 如果函数  $F$  的导数存在且在  $h'$  的极小区域内连续, 则有  $F'(h') = 0$ , 因此通过迭代应用式(4)的更新规则可以获得让  $F(h)$  收敛至局部最小值的  $h$  取值系列, 即有

$$F(h^{\min}) \leq \dots \leq F(h^{t+1}) \leq F(h^t) \leq \dots \leq F(h^2) \leq F(h^1) \leq F(h^0).$$

### 2.2 基于 NMF 的社区挖掘

NMF 已被证明与  $K$ -means 聚类以及谱聚类方法紧密相关<sup>[26]</sup>, 并广泛应用于图像分析<sup>[27]</sup>、文档聚类<sup>[28]</sup>以及协同过滤<sup>[29]</sup>等研究领域。在复杂网络社区

挖掘研究领域,由于复杂网络适合采用图结构进行表达,而矩阵又是图结构数据常用的存储形式,因此基于 NMF 的社区挖掘方法已吸引大量关注,具有代表性的工作分别有:Wang 等人<sup>[30]</sup>探讨如何应用 NMF 对有向、无向以及混合网络进行社区挖掘的方法,并给出了相关迭代优化求解算法以及实验验证结果;Psorakis 等人<sup>[31]</sup>则提出一种基于贝叶斯 NMF 模型的社区挖掘方法,不仅能够挖掘重叠社区,而且具有能够自动确定社区挖掘数量的优势,但需要学习多个概率模型参数;Zhang 等人<sup>[32]</sup>提出一种带约束的 NMF 三分解模型 BNMTF,能够有效建模和发现节点的社区成员关系以及社区间的交互程度,并能够应用统一的模型发现有向网络或无向网络中的重叠社区;Lin 等人<sup>[33]</sup>针对社交网络用户具有多关系网络的社区挖掘问题,提出一种基于 NMF 的社区挖掘框架 MetaFac,可以使用关系超图统一建模社交网络用户多个关系网络并应用基本 NMF 模型进行社区挖掘;Nguyen 等人<sup>[34]</sup>关注重叠社区的挖掘,基于 NMF 分别设计了有向以及无向网络的社区挖掘方法,并首次采用 Kullback-leibler 离散度作为目标函数;Zhang 等人<sup>[35]</sup>提出一种应用字典学习算法改进的 NMF 社区挖掘模型,实验证明该模型具有运算时间复杂度低以及可有效对大规模复杂网络进行社区挖掘的优势;此外,Zhang 等人还在文献<sup>[36]</sup>中讨论一种基于对称二值 NMF 的重叠社区挖掘方法,可以为重叠社区的挖掘结果提供更直观的解释.虽然以上基于 NMF 的社区挖掘方法均能够有效挖掘社区结构,但均只考虑复杂网络节点链接信息进行社区挖掘,而忽视了属性信息对于提高社区挖掘质量的作用,同样无法保证社区成员具有相同属性.

### 2.3 集成链接和属性信息的社区挖掘

针对集成节点链接和属性信息进行社区挖掘的问题,Cheng 等人<sup>[19]</sup>提出一种基于图聚类算法的社区挖掘方法 SACluster,该方法首先通过设计统一距离度量函数进行图节点结构相似性和属性信息相似性的计算,然后采用 K-means 聚类算法把具有属性信息的复杂网络图划分为  $k$  个聚簇,每一聚簇都包含联系密集的子图以及子图的所有节点都具有同类属性. SACluster 虽然能够自动学习节点结构相似性和属性信息相似性对社区挖掘结果的贡献度,但整个社区挖掘过程需划分为多个阶段,需要再借助聚类算法挖掘社区,而且节点综合相似度的计算非常耗时,时间复杂度为  $O(n^3)$  ( $n$  为节点数),实际

应用中并不方便. Xu 等人<sup>[20]</sup>提出一种基于贝叶斯概率模型的属性图聚类方法 BAGC,该方法不用人工设计统一距离度量函数计算节点链接结构相似性和属性相似性,而是将节点聚类转化为概率推理问题,虽然在一定程度上简化了社区挖掘过程,但该方法需要学习确定多个模型参数进行概率推理. Ruan 等人<sup>[21]</sup>针对网络节点链接信息存在噪音影响社区挖掘质量的问题,设计一种通过融合链接强度和节点内容信息相似度增强两个节点间信号的方法,并通过偏置边抽样过程可以保留与图节点相关的边,最后通过 Metis 以及 Markov 聚类算法对最终形成的主干图进行社区挖掘,但该方法与 SACluster 方法一样也存在多阶段进行社区挖掘的问题. 本文工作与以上相关工作相比的创新之处在于设计了一种基于 NMF 可统一集成节点链接和属性信息的社区挖掘方法,挖掘过程更加直接有效,社区挖掘结果可保证成员节点链接更加紧密且属性更加相同,从而保证挖掘的社区更有意义和应用价值.

## 3 LANMF:集成链接和属性信息的社区发现方法

### 3.1 问题描述

复杂网络适合采用图结构形式进行表达,对于一个既包含链接信息又包含属性信息的复杂网络可以表示为  $G=(V,E,A)$ ,其中  $V=\{v_1,v_2,\dots,v_n\}$  为节点集合,节点数量  $n=|V|$ ,  $E$  为链接边集合,  $A=\{a_1,a_2,\dots,a_m\}$  为节点属性集合,属性数量  $m=|A|$ . 对于每一个节点  $v_i \in V$ ,都关联一个  $m \times 1$  维的属性向量  $[y_{1i}, y_{2i}, \dots, y_{mi}]^T$ ,其中  $y_{ji}$  表示节点  $v_i$  在属性  $a_j$  上的取值,且  $y_{ji} \geq 0$ ,所有节点的属性向量可以构成属性关联矩阵  $\mathbf{Y} \in \mathbb{R}_+^{m \times n}$ . 此外,不失一般性,图  $G$  可以建模为无向无权图,其对应的链接矩阵可以表示为  $\mathbf{X} \in \mathbb{R}_+^{n \times n}$ ,其中当节点  $i$  和  $j$  存在链接时,  $x_{ij} = x_{ji} = 1$ ,否则  $x_{ij} = x_{ji} = 0$ ,因此可知  $\mathbf{X}$  为对称矩阵. 设置  $\mathbf{C}=\{c_1, c_2, \dots, c_k\}$  为待挖掘的社区集合,  $k$  为指定的社区挖掘数,  $\mathbf{H} \in \mathbb{R}_+^{n \times k}$  用于表示节点与社区的归属关系矩阵,  $h_{ij} \geq 0$ ,那么本文社区挖掘问题可以描述为在给定复杂网络链接矩阵  $\mathbf{X}$  和属性关联矩阵  $\mathbf{Y}$  的条件下,基于 NMF 得到分解矩阵  $\mathbf{H}$ ,并根据  $\mathbf{H}$  计算各社区所包含的节点集合.

### 3.2 链接矩阵和属性关联矩阵联合分解模型

在问题描述中,  $\mathbf{X}$  和  $\mathbf{Y}$  均为非负矩阵,为保证社区节点链接紧密和属性更加相同,可以对  $\mathbf{X}$  和  $\mathbf{Y}$  进

行联合分解,并且都要求具有共同分解因子矩阵  $\mathbf{H}$ . 实际上  $\mathbf{X}$  和  $\mathbf{Y}$  都可以单独分解,如  $\mathbf{X}$  可以转换为社区挖掘的三分解(Tri-Factorization)形式  $\mathbf{X} \approx \mathbf{H}\mathbf{S}\mathbf{H}^T$ , 其中  $\mathbf{H}$  为社区归属关系矩阵,  $\mathbf{S}$  为社区连接强度矩阵. 由于  $\mathbf{X}$  为对称矩阵,所以  $\mathbf{S}$  也为对称矩阵,令  $\mathbf{H} \leftarrow \mathbf{H}\mathbf{S}^{1/2}$ , 那么基于  $\mathbf{X}$  的矩阵三分解形式可以进一步简化为  $\mathbf{X} \approx \mathbf{H}\mathbf{H}^T$ , 其中  $\mathbf{H}$  仍可以表示为社区归属关系矩阵. 对于属性关联矩阵  $\mathbf{Y}$ , 每一个列向量对应一个节点的属性分布向量,所以  $\mathbf{Y}$  类似于词项-文档矩阵,可以借鉴文献[28]基于 NMF 的文档聚类思想,将  $\mathbf{Y}$  分解为 NMF 的标准形式  $\mathbf{Y} \approx \mathbf{W}\mathbf{H}^T$ , 其中  $\mathbf{H}$  表示的含义同上,  $\mathbf{W} \in \mathbb{R}^{m \times k}$  用于表示属性与社区之间的关联强度矩阵. 综合以上  $\mathbf{X}$  和  $\mathbf{Y}$  的独立分解形式,采用 Frobenius 范数进行误差度量,得到两者的联合分解模型为

$$\min_{\mathbf{H} \geq 0, \mathbf{W} \geq 0} \{F(\mathbf{H}, \mathbf{W}) = \frac{1}{2}(\|\mathbf{X} - \mathbf{H}\mathbf{H}^T\|_F^2 + \|\mathbf{Y} - \mathbf{W}\mathbf{H}^T\|_F^2 + \lambda \|\mathbf{H}\|_F^2 + \varphi \|\mathbf{W}\|_F^2)\} \quad (5)$$

其中  $\|\mathbf{H}\|_F^2$  和  $\|\mathbf{W}\|_F^2$  为正则化项,用于提高模型的稳定性,  $\lambda$  和  $\varphi$  为正则化参数,系数 1/2 用于简化后续运算. 根据矩阵 Trace 与 Frobenius 范数的关系,式(5)中的目标函数  $F$  可以重写为

$$F = \frac{1}{2}(tr(\mathbf{X}\mathbf{X}^T) - 2tr(\mathbf{X}\mathbf{H}\mathbf{H}^T) + tr(\mathbf{H}\mathbf{H}^T\mathbf{H}\mathbf{H}^T) + tr(\mathbf{Y}\mathbf{Y}^T) - 2tr(\mathbf{Y}\mathbf{H}\mathbf{W}^T) + tr(\mathbf{W}\mathbf{H}^T\mathbf{H}\mathbf{W}^T) + \lambda tr(\mathbf{H}^T\mathbf{H}) + \varphi tr(\mathbf{W}^T\mathbf{W})) \quad (6)$$

最小化目标函数  $F$  可获得  $\mathbf{H}$  和  $\mathbf{W}$  的近似分解结果,由于  $\forall h_{ij} \in \mathbf{H}$  均有  $h_{ij} \geq 0$ ,  $\forall \omega_{pq} \in \mathbf{W}$  均有  $\omega_{pq} \geq 0$ , 最小化  $F$  可以转化为典型的受限约束求极值问题,可以应用拉格朗日乘数方法进行求解. 设  $\alpha_{ij}$  和  $\beta_{pq}$  分别为受限条件  $h_{ij} \geq 0$  以及  $\omega_{pq} \geq 0$  对应的拉格朗日乘数,且  $\alpha = [\alpha_{ij}]$ ,  $\beta = [\beta_{pq}]$ , 则  $F$  对应的拉格朗日乘数函数  $L$  为  $L = F + tr(\alpha\mathbf{H}^T) + tr(\beta\mathbf{W}^T)$ . 为优化求解函数  $L$ , 可以引入 Karush-Kuhn-Tucker(KKT) 条件,有

$$\frac{\partial L}{\partial \mathbf{H}} = -2\mathbf{X}\mathbf{H} + 2\mathbf{H}\mathbf{H}^T\mathbf{H} - \mathbf{Y}^T\mathbf{W} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H} + \alpha = 0 \quad (7)$$

$$\frac{\partial L}{\partial \mathbf{W}} = -\mathbf{Y}\mathbf{H} + \mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W} + \beta = 0 \quad (8)$$

另外,根据 KKT 的平滑条件,有  $\alpha_{ij}h_{ij} = 0$  及  $\beta_{pq}\omega_{pq} = 0$ , 则由式(7)和(8)可分别得到如下新的等式:

$$(2\mathbf{X}\mathbf{H} + \mathbf{Y}^T\mathbf{W})_{ij}h_{ij} - (2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij}h_{ij} = 0 \quad (9)$$

$$(\mathbf{Y}\mathbf{H})_{pq}\omega_{pq} - (\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}\omega_{pq} = 0 \quad (10)$$

由式(9)和(10)可分别得到  $h_{ij}$  和  $\omega_{pq}$  的乘性迭代更新规则为

$$h_{ij} = \frac{(2\mathbf{X}\mathbf{H} + \mathbf{Y}^T\mathbf{W})_{ij}}{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij}}h_{ij} \quad (11)$$

$$\omega_{pq} = \frac{(\mathbf{Y}\mathbf{H})_{pq}}{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}\omega_{pq} \quad (12)$$

式(11)和(12)为目标函数  $F$  的受限优化求解规则,需要证明应用这些迭代规则可以保证目标函数  $F$  为非增函数,即  $F$  可以不断收敛获得极小值,本文的具体证明过程见附录 1.

### 3.3 社区挖掘算法

在应用迭代更新规则分别获得  $\mathbf{H}$  和  $\mathbf{W}$  的局部最优解后,节点的社区归属关系可以通过矩阵  $\mathbf{H}$  直接判断.  $\forall h_{ij} \in \mathbf{H}$  表示节点与社区之间的归属关系强度,那么根据  $h_{ij}$  的大小可以确定节点  $v_i$  的归属社区,具体社区挖掘算法描述如算法 1.

**算法 1.** 集成链接和属性信息的社区挖掘算法.

输入:  $\mathbf{X}, \mathbf{Y}, V, k, \lambda, \varphi$ , 迭代次数  $t$

输出:  $C = \{c_1, c_2, \dots, c_k\}$

步骤:

1. 随机初始化非负值矩阵  $\mathbf{H}$  和  $\mathbf{W}$ ;
2.  $iter \leftarrow 0$ ;
3. WHILE ( $iter < t$ )
4. 分别应用式(11)和(12)更新  $\forall h_{ij} \in \mathbf{H}$  以及  $\forall \omega_{pq} \in \mathbf{W}$ ;
5.  $iter = iter + 1$ ;
6. END WHILE
7.  $\forall i = 1, 2, \dots, k, c_i \leftarrow \emptyset$ ;
8. FOR EACH  $v_i \in V$  DO
9.  $p = \arg \max_i h_{it}$ ;
10.  $c_p = c_p \cup \{v_i\}$ ;
11. END FOR

算法 1 中迭代应用式(11)和(12)更新  $\mathbf{H}$  和  $\mathbf{W}$  的元素是最耗时的步骤,并且都集中在矩阵相乘的操作中. 由于现实中的大规模复杂网络对应的  $\mathbf{X}$  和  $\mathbf{Y}$  均为稀疏矩阵,所以可以采用稀疏矩阵的优化乘法. 假设  $\mathbf{X}$  和  $\mathbf{Y}$  的非 0 元素个数分别为  $\mu_X$  和  $\mu_Y$ , 那么可以计算算法 1 的时间复杂度为  $O(t(\mu_X k + \mu_Y k + nk^2))$ . 随着节点数  $n$  的增多,单个节点具有的链接数以及属性数远小于  $n$ , 近似相当于常数项,那么可推算  $\mu_X = O(n)$  以及  $\mu_Y = O(n)$ , 因此算法 1 具有跟  $n$  线性相关的时间复杂度.

现实中的社区之间可能会存在重叠的情况,实际表现为重叠节点在重叠社区中的社区归属强度值都比较接近. 为更好地度量这种重叠关系,可以根据

算法 1 得到的  $\mathbf{H}$  矩阵计算每一个节点与各社区的绝对归属感 (Absolute Membership Degree, AMD). 本文对任意一个节点  $v_i$  与社区  $j$  的绝对归属感  $AMD_{ij}$  定义为

$$AMD_{ij} = \frac{h_{ij} - \min_i h_{ij}}{\max_i h_{ij} - \min_i h_{ij}} \quad (13)$$

其中  $\min_i h_{ij}$  和  $\max_i h_{ij}$  分别表示节点  $v_i$  的社区归属强度向量  $[h_{i1}, h_{i2}, \dots, h_{ik}]$  中的最小分量和最大分量, 因此可知  $AMD_{ij} \in [0, 1]$ . 通过设置阈值  $\epsilon$ , 对于任意节点  $v_i$ , 当  $AMD_{ij} > \epsilon$  时,  $v_i$  划入社区  $j$ , 当  $v_i$  的多个社区绝对归属感都大于  $\epsilon$  时, 则  $v_i$  可以同时划入多个社区, 从而可以得到重叠社区挖掘结果.

算法 1 基于矩阵  $\mathbf{H}$  挖掘社区, 而  $\mathbf{W}$  为属性与社区之间的关联强度矩阵, 可以用于挖掘各社区的强关联属性. 如应用算法 1 对图 1 所示的社交网络示例进行社区挖掘, 为方便同时验证重叠社区的挖掘策略, 设置  $k=3$ , 属性集合  $A=\{\text{大数据, 社交网络}\}$ ,  $\lambda$  和  $\varphi$  均取值为 0.5, 当目标函数迭代收敛时, 可得到  $\mathbf{H}$  和  $\mathbf{W}$  的近似分解结果分别为

$$\mathbf{H}^T = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.34 & 0.35 & 0.42 & 0.34 & 0.00 & 0.00 & 0.00 & 0.46 & 0.46 & 0.46 \\ 0.07 & 0.05 & 0.41 & 0.07 & 0.64 & 0.59 & 0.59 & 0.00 & 0.00 & 0.00 \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} 0.47 & 2.39 & 0.00 \\ 2.07 & 0.11 & 0.00 \end{bmatrix}.$$

根据  $\mathbf{H}$  可以得到社区挖掘结果为  $c_1=\{4, 5, 6\}$ ,  $c_2=\{0, 1, 2, 3, 7, 8, 9\}$ ,  $c_3$  为空, 这与图 4 所示的集成链接和属性信息的社区挖掘结果一致. 值得注意的是, 节点 2 归属  $c_1$  和  $c_2$  的强度值较接近, 分别为 0.41 和 0.42, 可计算其与  $c_1$  和  $c_2$  的绝对归属感分别为 0.98 和 1. 容易判断  $c_1$  和  $c_2$  是重叠社区, 重叠节点集合为  $\{2\}$ . 另外, 从  $\mathbf{W}$  可以看出属性“大数据”与社区  $c_2$  的关联强度为 2.39, 远大于属性“社交网络”与社区  $c_2$  的关联强度 0.11, 可以理解为社区  $c_2$  所关注的研究兴趣主题为“大数据”, 这跟  $c_2$  中的节点都具有“大数据”属性是密切相关的, 同样可以分析得到  $c_1$  所关注的研究兴趣主题为“社交网络”. 以上对 OSN 示例的社区挖掘应用分析表明, 算法 1 可以基于联合 NMF 统一集成节点的链接和属性信息挖掘社区结构, 而且可以同时挖掘各社区所关联的兴趣主题属性信息. 此外, 还具有“模糊聚类”的特点, 可以通过计算节点的绝对归属感挖掘重叠社区及重叠节点.

## 4 实验及应用分析

### 4.1 实验数据集

为验证 LANMF 方法的有效性, 选择了 6 个含有链接和属性信息的典型复杂网络数据集进行实验, 各数据集的具体信息分别描述如下:

(1) Political Blogs 数据集<sup>[19]</sup>. 该数据集包含 1490 个节点和 19090 条边, 每一个节点代表一个有关美国政治的博客网页, 每一条边代表网页间的超链接关系, 由于 LANMF 考虑建模复杂网络为无向图的情况, 因此忽略了超链接的有向性, 最后剩余 16715 条边. 每一个节点都关联一个表明博客网页政治倾向的属性, 取值为 liberal 或者 conservative.

(2) CiteSeer 数据集<sup>[37]</sup>. 该数据集包含 3312 个节点和 4732 条边, 每一个节点代表一篇科技文献, 每一条边代表科技文献之间的引用关系, 此外每一个节点都关联一个类别属性, 总的类别属性取值数目为 6.

(3) Cora 数据集<sup>[37]</sup>. 该数据集同 CiteSeer 一样同为科技文献引文网络数据集, 包含 2708 个节点和 5429 条边, 每一个节点都关联一个类别属性, 总的类别属性取值数目为 7.

(4) WebKB 数据集<sup>[37]</sup>. 该数据集由 Cornell、Texas、Washington 以及 Wisconsin 等 4 所大学的网页以及网页间的链接关系数据组成, 共包含 874 个节点和 1608 条边, 每一个节点都关联一个类别属性, 总的类别属性取值数目为 5.

(5) 中国计算机学会中文会刊合著关系网络数据集 CJCA. 开发 Web 爬虫程序抽取了中国计算机学会中文会刊, 如《计算机学报》、《软件学报》以及《计算机研究与发展》等 12 种期刊近 5 年发表的文章数据构建合著关系网络数据集, 该数据集包含 49453 个节点和 117146 条边, 每一个节点代表一个作者, 每一条边代表两个作者存在的合著关系. 此外, 使用每一个作者所发表的文章关键词作为其对应节点所关联的属性, 总的属性取值数目为 70632.

(6) 学者网社交网络数据集<sup>[38]</sup>. 学者网提供个人学术信息管理以及主页服务, 并且为用户提供实名制社交网络服务, 用户之间可以建立好友关系并形成好友关系网络. 该数据集抽取了学者网中研究领域为社交网络或无线传感器网络的用户公开的好友关系数据, 共包含 101 个用户节点以及 948 条边, 每一个用户节点都关联一个研究兴趣属性: 社交

网络或无线传感器网络。

表 1 给出了各数据集的特征信息,其中 Political Blogs、CiteSeer、Cora 以及 WebKB 数据集均用于 LANMF 与现有代表性社区挖掘方法进行对比分析,CJCA 与学者网数据集则作为具体应用实例,用于应用 LANMF 进行研究兴趣主题社区挖掘以及重叠社区挖掘,并分析社区挖掘效果.实验采用的编程语言为 Java,运行机器具有主频为 2.93 GHz 的处理器,8GB 内存以及 1 TB 的磁盘存储空间。

| 表 1 数据集特征       |        |         |        |
|-----------------|--------|---------|--------|
| 数据集             | V      | E       | A      |
| Political Blogs | 1490   | 16 715  | 2      |
| CiteSeer        | 3312   | 4732    | 6      |
| Cora            | 2708   | 5429    | 7      |
| WebKB           | 874    | 1608    | 5      |
| CJCA            | 49 453 | 117 146 | 70 632 |
| 学者网             | 101    | 948     | 2      |

4.2 评价准则与比较方法

对于给定的社区挖掘结果  $C = \{c_1, c_2, \dots, c_k\}$ , 为评价集成链接和属性信息的社区挖掘质量,采用了文献[19]提出的两种评价准则:社区链接密度  $Density$  和社区属性信息熵  $Entropy$ ,分别定义如下:

$$Density(\{c_i\}_{i=1}^k) = \frac{\sum_{i=1}^k \frac{|\{(v_p, v_q) | v_p, v_q \in c_i, (v_p, v_q) \in E\}|}{|E|}}{k} \quad (14)$$

| 表 2 Political Blogs 数据集社区挖掘结果评价 |               |               |               |               |               |               |               |               |
|---------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 算法                              | k=3           |               | k=5           |               | k=7           |               | k=9           |               |
|                                 | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       |
| CDAgent                         | 0.6754        | 0.9012        | 0.6635        | 0.9146        | 0.6218        | 0.7864        | 0.6147        | 0.7649        |
| BNMTF                           | 0.6867        | 0.9940        | 0.6880        | 0.8972        | 0.6402        | 0.8861        | 0.6257        | 0.9215        |
| SACluster                       | 0.8526        | 0.6413        | 0.7762        | 0.4465        | 0.6917        | 0.2024        | 0.6357        | 0.3746        |
| BAGC                            | 0.7087        | 0.6023        | 0.7564        | 0.3901        | 0.7189        | 0.4756        | 0.7668        | 0.3796        |
| LANMF                           | <b>0.9030</b> | <b>0.0145</b> | <b>0.8735</b> | <b>0.0231</b> | <b>0.8643</b> | <b>0.0487</b> | <b>0.8419</b> | <b>0.0576</b> |

| 表 3 CiteSeer 数据集社区挖掘结果评价 |               |               |               |               |               |               |               |               |
|--------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 算法                       | k=6           |               | k=12          |               | k=15          |               | k=20          |               |
|                          | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       |
| CDAgent                  | 0.2864        | 6.3547        | 0.2914        | 6.2574        | 0.2319        | 7.1576        | 0.1175        | 7.9846        |
| BNMTF                    | 0.2763        | 6.9586        | 0.2984        | 6.4267        | 0.2420        | 7.7462        | 0.1243        | 8.2765        |
| SACluster                | 0.3975        | 5.4531        | 0.3012        | 6.1638        | 0.2746        | 6.8756        | 0.1796        | 6.7841        |
| BAGC                     | 0.3486        | 5.9746        | 0.3968        | 5.4756        | 0.2476        | 6.4756        | 0.1591        | 5.7468        |
| LANMF                    | <b>0.4095</b> | <b>2.5629</b> | <b>0.4328</b> | <b>2.4251</b> | <b>0.3922</b> | <b>2.5629</b> | <b>0.4123</b> | <b>2.7629</b> |

| 表 4 Cora 数据集社区挖掘结果评价 |               |               |               |               |               |               |               |               |
|----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 算法                   | k=7           |               | k=10          |               | k=15          |               | k=20          |               |
|                      | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       | Density       | Entropy       |
| CDAgent              | 0.2364        | 7.3479        | 0.2247        | 6.9976        | 0.2533        | 6.9971        | 0.2347        | 8.6423        |
| BNMTF                | 0.1867        | 7.4766        | 0.2174        | 7.4518        | 0.2846        | 7.4189        | 0.2107        | 8.5245        |
| SACluster            | 0.2674        | 6.4852        | 0.2886        | 5.4686        | 0.3108        | 6.9745        | 0.2707        | 7.4821        |
| BAGC                 | 0.2984        | 7.2151        | 0.2543        | 6.8476        | 0.3415        | 6.0412        | 0.2469        | 8.1459        |
| LANMF                | <b>0.3985</b> | <b>2.6237</b> | <b>0.4124</b> | <b>2.6225</b> | <b>0.4534</b> | <b>2.6050</b> | <b>0.4368</b> | <b>2.6191</b> |

$$Entropy(\{c_i\}_{i=1}^k) = \sum_{i=1}^m \sum_{j=1}^k \frac{|c_j|}{|V|} entropy(a_i, c_j) \quad (15)$$

其中  $entropy(a_i, c_j) = -p_{ij} \log_2 p_{ij}$ ,  $p_{ij}$  为社区  $j$  中的节点具有属性值  $a_i$  的比例.理想的社区挖掘结果不仅要求社区节点链接紧密,而且要求节点属性高度相同,所以  $Density$  值越大,并且  $Entropy$  值越小,社区挖掘的结果质量就越好.为对比验证 LANMF 的优势,选择了 4 种目前具有代表性的社区挖掘方法,包括文献[18]提出的基于多 Agent 的无符号网络社区挖掘方法 CDAgent、文献[19]提出的 SACluster、文献[20]提出的 BAGC 方法以及文献[32]提出的基于 NMF 的社区挖掘方法 BNMTF,其中 CDAgent 和 BNMTF 只考虑了节点链接信息,而 SACluster 和 BAGC 可同时利用节点链接信息和属性信息。

4.3 实验对比结果分析

表 2~表 4 分别列出 4 种方法在 Political Blogs、CiteSeer、Cora 以及 WebKB 数据集上的社区挖掘评价结果,其中 LANMF 方法中的参数  $\lambda$  和  $\varphi$  均设置为 0.5.对于属性取值数目为 2 的 Political Blogs 数据集,分别设置社区挖掘数  $k=3,5,7,9$ ,从表 2 可以看出对于只考虑链接信息进行社区挖掘的方法 CDAgent 和 BNMTF,由于没有集成利用属性信息,虽然随着  $k$  值的增加,  $Density$  值都稳定在 0.6

表 5 WebKB 数据集社区挖掘结果评价

| 算法        | $k=5$          |                | $k=8$          |                | $k=15$         |                | $k=20$         |                |
|-----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|           | <i>Density</i> | <i>Entropy</i> | <i>Density</i> | <i>Entropy</i> | <i>Density</i> | <i>Entropy</i> | <i>Density</i> | <i>Entropy</i> |
| CDAgent   | 0.2615         | 6.1456         | 0.2947         | 7.1245         | 0.3154         | 7.0125         | 0.2287         | 7.8642         |
| BNMTF     | 0.2587         | 6.5476         | 0.2898         | 7.5478         | 0.3078         | 7.2419         | 0.2325         | 8.6412         |
| SACluster | 0.2985         | 5.2314         | 0.3134         | 5.8961         | 0.3756         | 6.4723         | 0.2674         | 6.4237         |
| BAGC      | 0.2743         | 3.2587         | 0.3013         | 4.1247         | 0.3563         | 5.1479         | 0.2612         | 6.8419         |
| LANMF     | <b>0.3849</b>  | <b>1.6232</b>  | <b>0.4286</b>  | <b>1.5820</b>  | <b>0.4017</b>  | <b>1.5806</b>  | <b>0.3853</b>  | <b>1.5430</b>  |

以上, 但 *Entropy* 值都相对比较大, 表明 CDAgent 和 BNMTF 方法所挖掘的社区各成员节点存在较大的属性取值差异, 社区挖掘的质量不高. 而对于可结合链接信息和属性信息进行社区挖掘的 3 种方法 SACluster、BAGC 和 LANMF, *Density* 值和 *Entropy* 值都优于 CDAgent 和 BNMTF 方法. 需要指出的是 LANMF 方法随着  $k$  值的增加, *Density* 值均大于 0.8 且 *Entropy* 值均小于 0.06, 表明 LANMF 所挖掘的社区不仅内部节点链接紧密而且属性相同的比例非常高, 这好于其他两种同类方法 SACluster 和 BAGC. 另外, 对于属性取值数目相对大些的 CiteSeer、Cora 以及 WebKB 数据集, 随着  $k$  值的增加, LANMF 方法的两种评价指标均处于理想的范围, 而 SACluster 和 BAGC 方法均表现不够稳定, 评价指标结果差异较大, 这表明 LANMF 方法具有更好的稳定性和可靠性. 其原因在于, LANMF 采用了可统一集成处理链接和属性信息的联合矩阵分解模型, 而 SACluster 和 BAGC 均需要采用孤立模型分别处理, 无法保证社区成员链接和属性特征的平衡统一. 此外, 需要指出的是 LANMF 方法另外一个显著优势是可以通过近似矩阵分解结果  $\mathbf{H}$  直接识别节点与社区的归属关系, 不需要再借助其他聚类算法进行二次处理获得社区挖掘结果, 这比 SACluster 和 BAGC 方法更加直接有效, 如图 5~图 8

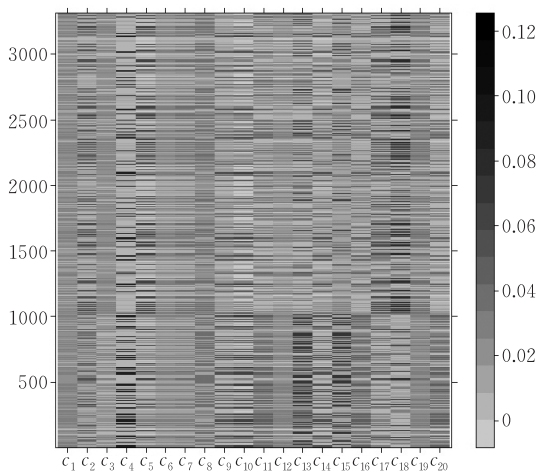


图 6 CiteSeer  $\mathbf{H}$  矩阵 HeatMap( $k=20$ )

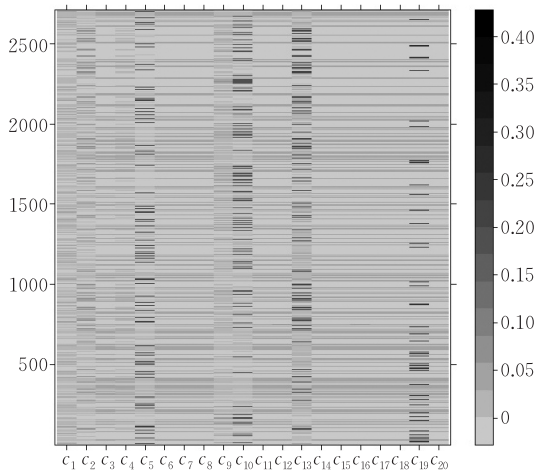


图 7 Cora  $\mathbf{H}$  矩阵 HeatMap( $k=20$ )

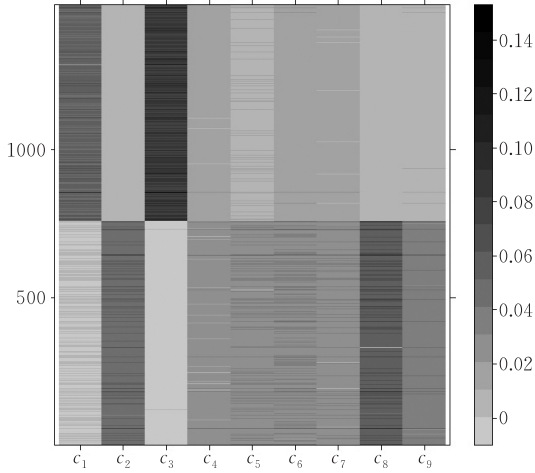


图 5 Political Blogs  $\mathbf{H}$  矩阵 HeatMap( $k=9$ )

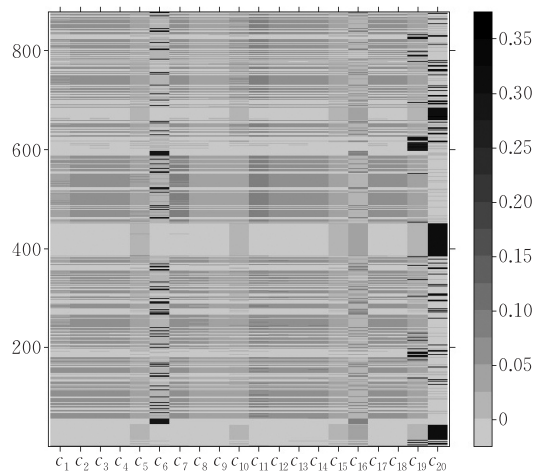


图 8 WebKB  $\mathbf{H}$  矩阵 HeatMap( $k=20$ )



分别为 LANMF 方法对以上 4 个数据集进行社区挖掘产生的  $\mathbf{H}$  矩阵对应的 HeatMap( $X$  轴代表社区  $ID$ ,  $Y$  轴代表节点  $ID$ ), 从中可以较清晰直观地通过  $\mathbf{H}$  矩阵各元素代表的色度强弱判断节点与社区的归属关系。

4.4 参数  $\lambda$  和  $\varphi$  取值分析

式(5)中的  $\lambda$  和  $\varphi$  为正则化参数, 用于控制求解规则的平滑性, 取值的大小将影响社区挖掘质量。为合理设置  $\lambda$  和  $\varphi$  的取值, 在 Political Blogs、CiteSeer、

Cora 以及 WebKB 数据集上进行了参数设置分析实验。考虑到正则化项  $\|\mathbf{H}\|_F^2$  和  $\|\mathbf{W}\|_F^2$  具有同等重要性, 可设置  $\lambda=\varphi=\theta$ , 并分析  $\theta$  在  $[0, 500]$  范围内取值对于各数据集社区挖掘质量的影响, 结果分别如图 9~图 12 所示。从结果可以看出  $\theta$  取值在  $[0, 1]$  范围内各数据集可获得稳定且较好的社区挖掘质量评价结果, 取值越大, 社区挖掘质量评价结果越差。根据该分析结果, 在本文的相关实验中,  $\lambda$  和  $\varphi$  均取值为 0.5。

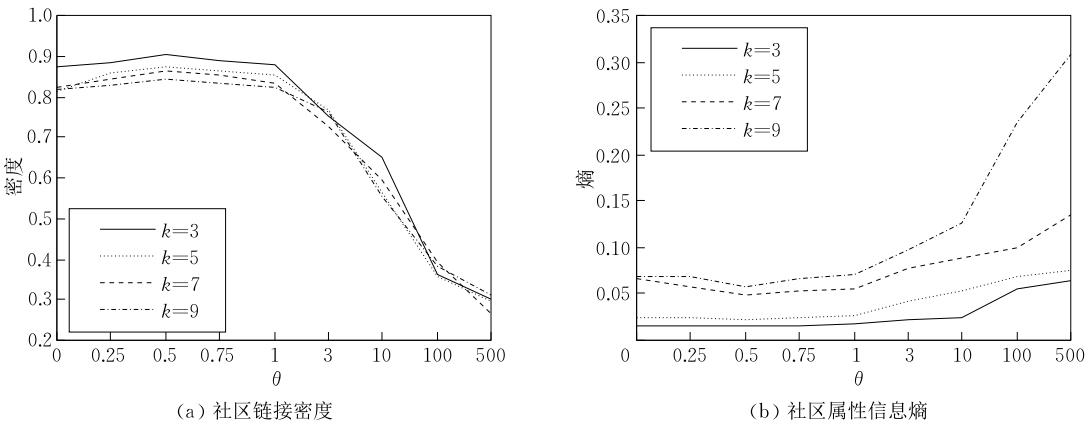


图 9 Political Blogs 不同  $\theta$  取值下的社区挖掘质量对比

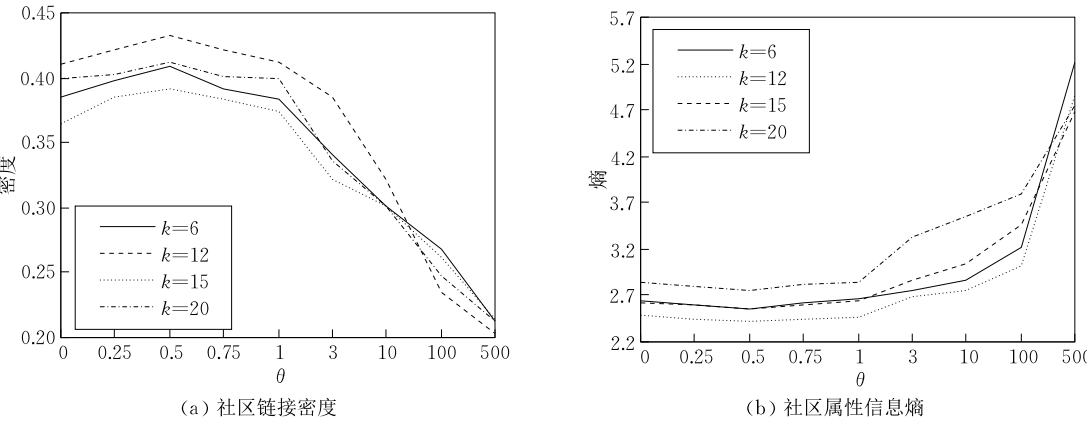


图 10 CiteSeer 不同  $\theta$  取值下的社区挖掘质量对比

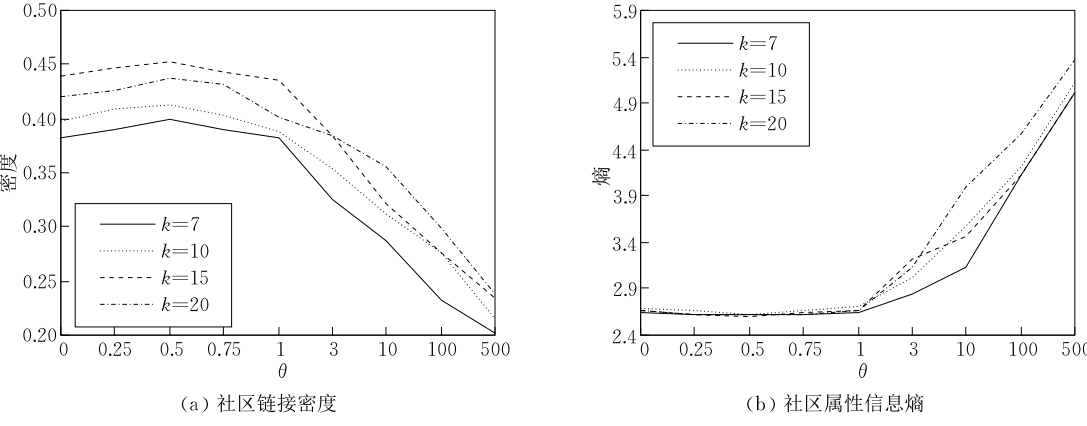


图 11 Cora 不同  $\theta$  取值下的社区挖掘质量对比

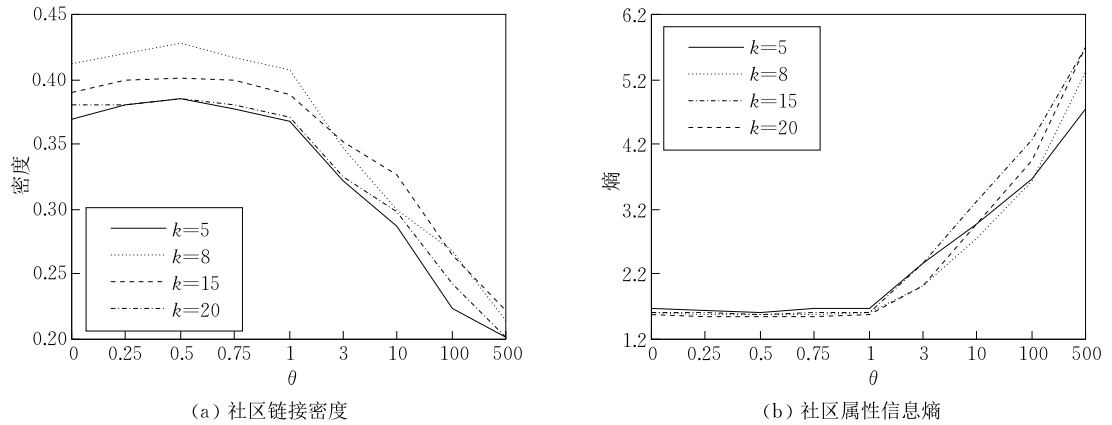


图 12 WebKB 不同  $\theta$  取值下的社区挖掘质量对比

4.5 应用实例

4.5.1 科技文献合著关系网络研究团队挖掘

科技文献合著关系的建立存在随意性,如存在非同一研究方向的作者合作发表论文的现象,或者即使是一个大研究团队的成员也存在研究方向不同,但合著发表论文的情况,所以如果单一利用合著关系信息挖掘科技文献合著关系网络中隐藏的研究兴趣主题社区,即研究团队,也将存在研究团队内作者研究兴趣属性差异较大的问题,所以有必要集成链接信息和属性信息进行挖掘. 本文应用 LANMF 方法对 CJCA 数据集进行研究团队的挖掘,在  $k=3000$  情况下,得到了各研究团队的成员以及关注的研究兴趣主题,表 6 列出了其中 6 个典型的研究团队所包含的主要成员以及各典型研究团队所关注的主要研究兴趣主题属性,其中研究团队成员姓名以及研究兴趣主题属性名称后面的数值分别代表成员与研究团队之间的归属强度以及研究兴趣主题属性与研究团队之间的关联强度,分别来自  $\mathbf{H}$  和  $\mathbf{W}$  矩阵. 成员与团队的归属强度越大,则表明该成员在该

团队中的重要性越强. 例如,云计算研究团队中的“罗军舟”以及闪存数据库研究团队中的“孟小峰”均在各自团队中担任“领袖”角色. 此外,与团队关联强度最大的兴趣主题属性则代表着团队所有成员的主要研究方向.

表 6 还可以有效体现 LANMF 同时集成节点链接和属性信息进行社区挖掘的效果,如“时空推理研究团队”和“复杂网络研究团队”的成员虽然同属于吉林大学符号计算与知识工程教育部重点实验室,两个团队之间的成员也存在交叉合著关系,例如,刘大有与“复杂网络研究团队”的成员也存在密切的合著关系. 如果应用 BNMTF 挖掘合著关系网络只得到 1 个研究团队,但 LANMF 的社区挖掘可以得到 2 个研究团队(图 13),两个团队内的成员研究兴趣主题属性更加相同:社区 1 的成员主要研究时空推理,社区 2 的成员则研究复杂网络,这表明 LANMF 可得到更准确的社区挖掘结果,得到的社区也更合理以及更具有可解释性.

表 6 CJCA 的典型研究团队

| 团队名称      | 核心成员                                                            | 主要研究兴趣主题属性                                                             |
|-----------|-----------------------------------------------------------------|------------------------------------------------------------------------|
| 云计算研究团队   | 罗军舟(0.4134); 宋爱波(0.3278); 东方(0.1968); 田田(0.1179); 朱 夏(0.1043)   | 云计算(1.1157); 网格(1.0859); 数据中心(0.9547); 绿色计算(0.8941); 对等网络(0.8286)      |
| 闪存数据库研究团队 | 孟小峰(0.6719); 张金增(0.1315); 赖文豫(0.1217); 范玉雷(0.1064); 马友忠(0.0981) | 闪存数据库(1.1787); 索引(0.9567); 查询处理(0.9466); 并行聚集(0.8994); 并行表扫描(0.8994)   |
| 信息安全研究团队  | 冯登国(0.5992); 张立武(0.1364); 宋海欣(0.0930); 刘凡保(0.0871); 谢 涛(0.0871) | 密钥恢复(1.0105); 密文策略(1.0088); 密钥协商(0.8592); 流密码算法(0.8373); 属性加密(0.8232); |
| 时空推理研究团队  | 刘大有(0.6176); 王生生(0.1490); 赖永(0.1467); 谷方明(0.1032); 吕倩楠(0.1032)  | 时空推理(1.0771); 代数封闭算法(1.0771); 模态逻辑(0.9142); Tableau 算法(0.9142)         |
| 复杂网络研究团队  | 杨博(0.4251); 金 弟(0.2402); 何东晓(0.2402); 黄 晶(0.1678); 刘杰(0.1410)   | 复杂网络(2.8196); 社区挖掘(1.6627); 网络聚类(1.6069); 社区结构(1.5200); 重叠社区挖掘(1.4659) |
| 移动互联网研究团队 | 王兴伟(0.5113); 黄敏(0.4336); 曲大鹏(0.2284); 孙佳佳(0.1349); 高程希(0.1349)  | 移动互联网(1.3801); 链路质量(1.2014); 移动对等网络(1.1928); 蚁群路由算法(1.1734)            |

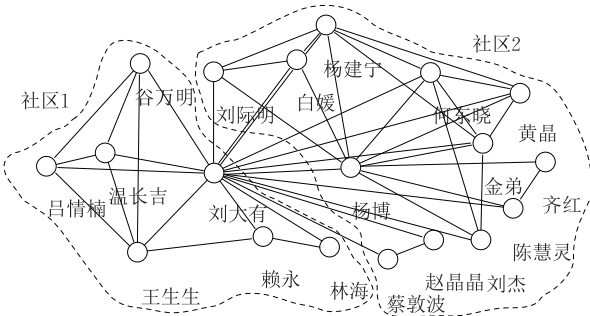


图 13 时空推理研究团队和复杂网络研究团队的合著关系网络社区挖掘

4.5.2 学者网社交网络研究兴趣主题社区挖掘

社交网络中好友关系的建立也存在随意性,具有好友关系的用户不一定都具有相同的兴趣.如果单一考虑用户好友链接关系信息挖掘社区,随意建立的好友链接将成为“噪音数据”,将降低社区挖掘的质量,所以对学者网社交网络进行研究兴趣主题社区挖掘,也有必要集成利用用户链接信息和属性信息.图 14 为学者网数据集对应的社交网络,其中研究兴趣属性为“社交网络”的节点都用圆形表示,研究兴趣属性为“无线传感器网络”的节点则用黑色方形表示,可以看出部分具有不同研究兴趣属性的成员节点之间也存在紧密的链接.假定社区挖掘数  $k=2$ ,应用只考虑链接信息的社区挖掘方法 BNMTF 挖掘社区,得到图 15 所示的社区挖掘结果,可以看出其中社区 2 的成员研究兴趣属性差异较大,结果评价指标分别为  $Density=0.6582$  和  $Entropy=0.8211$ . 另一方面,应用 LANMF 进行社区挖掘,得

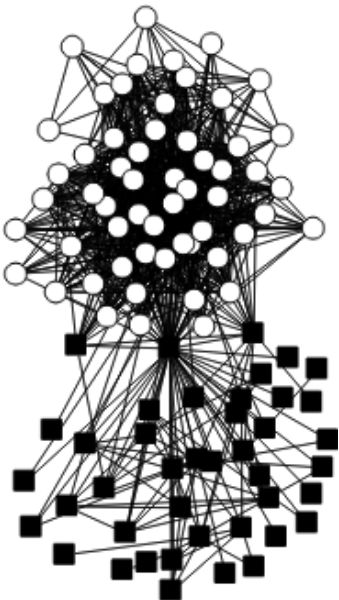


图 14 学者网社交网络

到图 16 所示的社区挖掘结果,可以看出社区 1 或社区 2 的成员研究兴趣属性都相同(表 7 列出了社区 1 和社区 2 的成员及各成员的社区归属强度),结果评价指标分别为  $Density=0.9209$  和  $Entropy=0.0$ ,可见 LANMF 社区挖掘的结果质量更好.这也进一步说明 LANMF 适合对现实中的复杂网络进行主题社区挖掘,获得的社区结果更有价值和意义.实际上 LANMF 通过联合 NMF 有效融合了用户节点链接和属性信息,两类信息互为补充,大大降低了“噪音数据”对社区挖掘质量的影响.

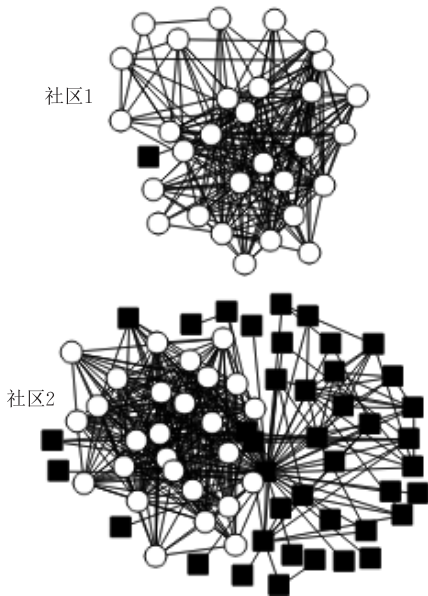


图 15 BNMTF 挖掘学者网社区

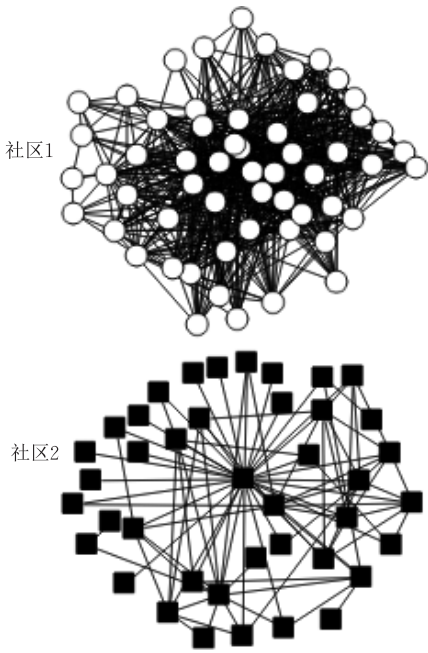


图 16 LANMF 挖掘学者网社区

表 7 学者网典型研究兴趣主题社区的主要成员

| 社交网络研究兴趣主题社区(社区 1)                                                                                                                                                                                                                                    | 无线传感器网络研究兴趣主题社区(社区 2)                                                                                                                                                                                                                                    |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 汤庸(0.2722);杨镇雄(0.2663);傅城州(0.2577);陈国华(0.2550);洪少文(0.2540);方家轩(0.2503);赖韩江(0.2356);蔡祖锐(0.2338);罗源(0.2316);李建国(0.2251);李春英(0.2241);陈洁敏(0.2238);蒋文丽(0.2235);贺超波(0.2222);袁成哲(0.2220);麦辉强(0.2215);张龙(0.2191);许玉赢(0.2186);朱佳(0.2103);刘晓(0.2054);林海(0.2030) | 舒磊(0.3444);陆建俭(0.2138);李祥杰(0.2019);朱旭锋(0.1947);刘明珠(0.1897);郑岸雄(0.1855);熊建斌(0.1853);张海铭(0.1833);林东城(0.1829);林馨蕾(0.1803);邱长江(0.1800);黄剑锋(0.1769);李沁雪(0.1746);刘泽许(0.1711);庞木兰(0.1698);杨忠明(0.1691);杨敏静(0.1634);李容(0.1600);叶奇明(0.1576);徐小玲(0.1547);张锋(0.1545) |

4.5.3 重叠社区的挖掘

从图 13 所示的科技文献合著关系网络可以看出,刘大有分别与社区 1 和社区 2 都具有较密切的联系.实际上,由于科研人员具有研究兴趣多样化的特点,可以同时与多个研究团队都有较高的绝对归属度,即合著关系网络存在重叠社区,这可以通过 3.3 节设计 的重叠社区挖掘策略进行处理.通过设置节点绝对归属度阈值  $\epsilon=0.6$ ,对 4.5.1 节的 CJCA 数据集社区挖掘结果进行后续处理,提取了 CJCA 中存在的重叠社区以及重叠节点,其中部分典型的重叠社区如表 8 所示.从表 8 可以看出,重叠节点的各社区绝对归属度并不是求和等于 1 的关系,而是相互独立的,这容易通过设置阈值对重叠社区进行挖掘.为评价重叠社区挖掘的质量,本文在保留式(14)和(15)表示的社区链接密度

$Density$  和属性信息熵  $Entropy$  评价准则的基础上,引入文献[39]提出的重叠社区结构模块度函数  $Q_{ov}$ ,其定义如下:

$$Q_{ov} = \frac{1}{2|E|} \sum_{i=1}^k \sum_{p,q \in c_i} \frac{1}{O_p O_q} \left( x_{pq} - \frac{k_p k_q}{2|E|} \right) \quad (16)$$

其中: $O_p$  和  $O_q$  分别表示节点  $v_p$  和  $v_q$  所在的社区数量; $k_p$  和  $k_q$  分别表示节点  $v_p$  和  $v_q$  的度.通过在  $[0,1]$  范围内设置不同的  $\epsilon$  值,提取 4.5.1 节的 CJCA 数据集社区挖掘结果的重叠社区,评价结果发现当  $\epsilon=0.6$  时, $Q_{ov}$ ,  $Density$  以及  $Entropy$  的值分别为 0.2674、0.3241 及 8.7834,均为最优值,表明在计算节点社区绝对归属度的基础上,通过设置合适的阈值  $\epsilon$  可以有效获得高质量的重叠社区.

表 8 CJCA 的典型重叠社区

| 社区名称    | 社区核心成员                                                      | 重叠节点 | AMD  |
|---------|-------------------------------------------------------------|------|------|
| 云计算     | 罗军舟(0.4134);宋爱波(0.3278);东方(0.1968);田田(0.1179);朱夏(0.1043)    | 罗军舟  | 1.00 |
| 移动通信网络  | 杨明(0.4823);罗军舟(0.3926);王维(0.2016);吴文甲(0.1771);何高峰(0.1675)   |      | 0.77 |
| 社交媒体挖掘  | 贾焰(0.4208);周斌(0.3909);丁兆云(0.2906);林旺群(0.2864);吴泉源(0.1217)   | 贾焰   | 1.00 |
| 网络安全    | 韩伟红(0.3811);邹鹏(0.3255);郑黎明(0.3005);贾焰(0.2593);李爱平(0.1892)   |      | 0.68 |
| 下一代互联网  | 林闯(0.4867);孟坤(0.3995);万剑雄(0.2978);向旭东(0.2978);苏文博(0.1479)   | 林闯   | 1.00 |
| 网络性能优化  | 林闯(0.3925);田立勤(0.2765);陈振国(0.2761);张琪(0.2637);沈苏彬(0.1554)   |      | 1.00 |
| 无线传感器网络 | 任倩倩(0.4046);李建中(0.2765);程思瑶(0.2372);王宇(0.2111);高宏(0.1254)   | 李建中  | 0.68 |
| 数据库     | 李建中(0.5763);刘显敏(0.3198);石胜飞(0.3044);宁钊(0.2111);王朝坤(0.1754)  |      | 1.00 |
| 传感器技术   | 李栋(0.4263);邱杰凡(0.3064);石海龙(0.3064);崔莉(0.2727);杜文振(0.1832)   | 崔莉   | 0.64 |
| 物联网     | 崔莉(0.4839);谢开斌(0.3678);陈海明(0.3564);黄庭培(0.3125);张招亮(0.2987)  |      | 1.00 |
| 密码学     | 熊金波(0.3935);李凤华(0.3222);姚志强(0.3222);马建峰(0.3125);刘西蒙(0.2116) | 马建峰  | 0.79 |
| 无线网络    | 马建峰(0.4395);姬文江(0.3604);马卓(0.3041);张俊伟(0.2941);田有亮(0.2868)  |      | 1.00 |

5 结束语

挖掘成员节点链接紧密且属性高度相同的社区更有意义和应用价值,本文所提出的 LANMF 社区挖掘方法基于非负矩阵联合近似分解模型,具有严格的数学理论基础,可统一集成节点链接和属性信息进行社区挖掘.大量实验和实际应用表明 LANMF 比现有代表性方法更加直接有效挖掘复杂网络中的高质量社区,而且能够保证社区成员链接紧密且属性高度相同,适合于挖掘现实世界复杂网络中的主题社区以及重叠社区.由于现实中的大规

模复杂网络节点数多,属性维度大,社区挖掘将涉及大数据量计算,这要求社区挖掘方法具有更高的计算效率.已有研究表明 NMF 适合采用分布式集群计算框架 MapReduce 实现,所以在下一阶段的研究工作中,将重点研究如何基于 MapReduce 进一步提高 LANMF 的可扩展能力,从而可以更好地解决大规模复杂网络的社区挖掘问题.

参 考 文 献

[1] Wang Li, Cheng Xue-Qi. Dynamic community in online social networks. Chinese Journal of Computers, 2015, 38(2): 219-237(in Chinese)

- (王莉, 程学旗. 在线社会网络的动态社区发现及演化. 计算机学报, 2015, 38(2): 219-237)
- [2] Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Science*, 2002, 99(12): 7821-7826
- [3] Lin You-Fang, Wang Tian-Yu, Tang Rui, et al. An effective model and algorithm for community detection in social networks. *Journal of Computer Research and Development*, 2012, 49(2): 337-345(in Chinese)
- (林友芳, 王天宇, 唐锐等. 一种有效的社会网络社区挖掘模型和算法. 计算机研究与发展, 2012, 49(2): 337-345)
- [4] Lin Wang-Qun, Lu Feng-Shun, Ding Zhao-Yun, et al. Parallel computing hierarchical community approach based on weighted-graph. *Journal of Software*, 2012, 23(6): 1517-1530 (in Chinese)
- (林旺群, 卢风顺, 丁兆云等. 基于带权图的层次化社区并行计算方法. 软件学报, 2012, 23(6): 1517-1530)
- [5] Jin Di, Liu Da-You, Yang Bo, et al. Fast Complex network clustering algorithm using local detection. *Acta Electronica Sinica*, 2011, 39(11): 2540-2546(in Chinese)
- (金弟, 刘大有, 杨博等. 基于局部探测的快速复杂网络聚类算法. 电子学报, 2011, 39(11): 2540-2546)
- [6] Shakarian P, Roos P, Callahan D, et al. Mining for geographically disperse communities in social networks by leveraging distance modularity//*Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago, USA, 2013: 1402-1409
- [7] Chai Bian-Fang, Yu Jian, Jia Cai-Yan, et al. Fast algorithm on stochastic block model for exploring general communities. *Journal of Software*, 2013, 24(11): 2699-2709(in Chinese)
- (柴变芳, 于剑, 贾彩燕等. 一种基于随机块模型的快速广义社区挖掘算法. 软件学报, 2013, 24(11): 2699-2709)
- [8] Leng Zuo-Fu. Community detection in complex networks based on greedy optimization. *Acta Electronica Sinica*, 2014, 42(4): 723-729(in Chinese)
- (冷作福. 基于贪婪优化技术的网络社区挖掘算法研究. 电子学报, 2014, 42(4): 723-729)
- [9] Costa G, Ortale R. Probabilistic analysis of communities and inner roles in networks: Bayesian generative models and approximate inference. *Social Network Analysis and Mining*, 2013, 3(4): 1015-1038
- [10] Hsieh C J, Tiwari M, Deepak A, et al. Organizational overlap on social networks and its applications//*Proceedings of the 22nd International Conference on World Wide Web*. Seoul, Korea, 2013: 571-582
- [11] Ball B, Karrer B, Newman M E J. An efficient and principled method for detecting communities in networks. *Physics Review E*, 2011, 84(3): 036103
- [12] Huang Fa-Liang, Huang Ming-Xuan, Yuan Chang-An, et al. Spectral clustering ensemble algorithm for discovering overlapping communities in social networks. *Control and Decision*, 2014, 29(4): 713-718(in Chinese)
- (黄发良, 黄名选, 元昌安等. 网络重叠社区挖掘的谱聚类集成算法. 控制与决策, 2014, 29(4): 713-718)
- [13] Zhu Mu, Meng Fan-Rong, Zhou Yong. Density-based link clustering algorithm for overlapping community detection. *Journal of Computer Research and Development*, 2013, 50(12): 2520-2530(in Chinese)
- (朱牧, 孟凡荣, 周勇. 基于链接密度聚类的重叠社区挖掘算法. 计算机研究与发展, 2013, 50(12): 2520-2530)
- [14] He D, Liu J, Liu D, et al. An ant-based algorithm with local optimization for community detection in large-scale networks. *Advances in Complex System*, 2012, 15(8): 1-26
- [15] Jin Di, Liu Jie, Yang Bo, et al. Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, 2011, 37(7): 873-882 (in Chinese)
- (金弟, 刘杰, 杨博等. 局部搜索与遗传算法结合的大规模复杂网络社区探测. 自动化学报, 2011, 37(7): 873-882)
- [16] Gong M, Fu B, Jiao L. Memetic algorithm for community detection in networks. *Physical Review E*, 2011, 84(5): 056101
- [17] Yang B, Liu J M, Liu D Y, et al. An autonomy-oriented computing approach to community mining in distributed and dynamic networks. *Autonomous Agents and Multi-Agent Systems*, 2010, 20(2): 123-157
- [18] Yang Yi. Detecting Communities from Social Network Based on Multi-Agent [M. S. dissertation]. Jilin University, Changchun, 2015(in Chinese)
- (杨懿. 基于多 Agent 的社会网络社区发现方法[硕士学位论文]. 吉林大学, 长春, 2015)
- [19] Cheng H, Zhou Y, Jeffrey X Y. Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data*, 2011, 5(2): 12
- [20] Xu Z Q, Ke Y P, Wang Y, et al. A model-based approach to attributed graph clustering//*Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. Scottsdale, USA, 2012: 505-516
- [21] Ruan Y Y, Fuchry D, Parthasarathy S. Efficient community detection in large networks using content and links//*Proceedings of the 22nd International Conference on World Wide Web*. Seoul, Korea, 2013: 1089-1098
- [22] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401(10): 788-791
- [23] Lee D D, Seung H S. Algorithms for nonnegative matrix factorization//*Proceedings of the 24th Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2000: 556-562
- [24] Liu H W, Li X L, Zheng X Y. Solving non-negative matrix factorization by alternating least squares with a modified strategy. *Data Mining and Knowledge Discovery*, 2013, 26(3): 435-451

- [25] Ma S Q, Goldfarb D, Chen L F. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 2011, 128(1): 321-353
- [26] Ding C, He X F, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering//*Proceedings of the 5th SIAM International Conference on Data Mining*. California, USA, 2005: 606-610
- [27] Long X Z, Lu H T, Peng Y, et al. Graph regularized discriminative non-negative matrix factorization for face recognition. *Multimedia Tools and Applications*, 2014, 72(3): 2679-2699
- [28] Xu W, Liu X, Gong Y H. Document clustering based on non-negative matrix factorization//*Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada, 2003: 267-273
- [29] Kannan R, Ishteva M, Park H. Bounded matrix factorization for recommender system. *Knowledge and Information Systems*, 2014, 39(3): 491-511
- [30] Wang F, Li T, Wang X, et al. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 2011, 22(3): 493-521
- [31] Psorakis I, Roberts S, Ebden M, et al. Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E*, 2011, 83(6): 1-9
- [32] Zhang Yu, Yeung D Y. Overlapping community detection via bounded nonnegative matrix tri-factorization//*Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*. Beijing, China, 2012: 606-614
- [33] Lin Y R, Sun J, Castro P, et al. Metafac: Community discovery via relational hypergraph factorization//*Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France, 2009: 527-536
- [34] Nguyen, N P, Thai M T. Finding overlapped communities in online social networks with nonnegative matrix factorization //*Proceedings of the 2012 Military Communications Conference*. Orlando, Florida, 2012: 1-6
- [35] Zhang Z Y. Community structure detection in social networks based on dictionary learning. *Science China Information Sciences*, 2013, 56(7): 1-12
- [36] Zhang Z Y, Wang Y, Ahn Y Y. Overlapping community detection in complex networks using symmetric binary matrix factorization. *Physical Review E*, 2013, 87(6): 062803
- [37] Yang T B, Jin R, Chi Y, et al. A Bayesian framework for community detection integrating content and link//*Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. Montreal, Canada, 2009: 615-622
- [38] He Chao-Bo, Tang Yong, Shen Yu-Li, et al. Method for community mining using nonnegative matrix factorization model. *Journal of Chinese Computer Systems*, 2014, 35(6): 1275-1280(in Chinese)  
(贺超波, 汤庸, 沈玉利等. 应用非负值矩阵分解模型的社区挖掘方法. *小型微型计算机系统*, 2014, 35(6): 1275-1280)
- [39] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009, (3): P03024

附录 1. 正文中迭代更新规则的收敛性证明过程.  
证明乘性迭代更新规则:

$$h_{ij} = \frac{(2\mathbf{X}\mathbf{H} + \mathbf{Y}^T\mathbf{W})_{ij}}{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij}} h_{ij} \quad (\text{A1})$$

$$\omega_{pq} = \frac{(\mathbf{Y}\mathbf{H})_{pq}}{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}} \omega_{pq} \quad (\text{A2})$$

可保证如下目标函数可以不断收敛获得极小值:

$$F(\mathbf{H}, \mathbf{W}) = \frac{1}{2}(\|\mathbf{X} - \mathbf{H}\mathbf{H}^T\|_F^2 + \|\mathbf{Y} - \mathbf{W}\mathbf{H}^T\|_F^2 + \lambda\|\mathbf{H}\|_F^2 + \varphi\|\mathbf{W}\|_F^2) \quad (\text{A3})$$

根据正文 2.1 节通过引入辅助函数进行证明的思路, 首先, 对于  $\forall h_{ij} \in \mathbf{H}$ , 用函数  $F_{h_{ij}}(h)$  表示  $F$  中跟  $h_{ij}$  相关的部分. 显然,  $F_{h_{ij}}(h)$  在  $h_{ij}$  的一阶导数就是目标函数  $F$  关于  $h_{ij}$  的一阶偏导数, 根据正文式(7)可知:

$$F'_{h_{ij}}(h) = \frac{\partial F}{\partial h_{ij}} = (-2\mathbf{X}\mathbf{H} + 2\mathbf{H}\mathbf{H}^T\mathbf{H} - \mathbf{Y}^T\mathbf{W} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} \quad (\text{A4})$$

根据  $F'_{h_{ij}}(h)$  可计算  $F_{h_{ij}}(h)$  在  $h_{ij}$  的二阶导数  $F''_{h_{ij}}(h)$  为

$$F''_{h_{ij}}(h) = \frac{\partial(-2\mathbf{X}\mathbf{H})_{ij}}{\partial h_{ij}} + \frac{\partial(2\mathbf{H}\mathbf{H}^T\mathbf{H})_{ij}}{\partial h_{ij}} + \frac{\partial(\mathbf{H}\mathbf{W}^T\mathbf{W})_{ij}}{\partial h_{ij}} + \lambda,$$

$$\begin{aligned} \text{由于 } \frac{\partial(-2\mathbf{X}\mathbf{H})_{ij}}{\partial h_{ij}} &= -2\mathbf{X}_{ii}, \frac{\partial(2\mathbf{H}\mathbf{H}^T\mathbf{H})_{ij}}{\partial h_{ij}} = 2((\mathbf{H}\mathbf{H}^T)_{ii} + h_{ij}^2 + \sum_k h_{kj}^2), \frac{\partial(\mathbf{H}\mathbf{W}^T\mathbf{W})_{ij}}{\partial h_{ij}} = (\mathbf{W}^T\mathbf{W})_{jj}, \text{ 则有} \\ F''_{h_{ij}}(h) &= -2\mathbf{X}_{ii} + 2((\mathbf{H}\mathbf{H}^T)_{ii} + h_{ij}^2 + \sum_k h_{kj}^2) + (\mathbf{W}^T\mathbf{W})_{jj} + \lambda \end{aligned} \quad (\text{A5})$$

同理, 根据  $F'_{h_{ij}}(h)$  可以依次求得  $F_{h_{ij}}(h)$  在  $h_{ij}$  的三阶导数  $F^{(3)}_{h_{ij}}(h)$  及四阶导数  $F^{(4)}_{h_{ij}}(h)$ , 分别为  $F^{(3)}_{h_{ij}}(h) = 12h_{ij}$  和  $F^{(4)}_{h_{ij}}(h) = 12$ , 那么  $F_{h_{ij}}(h)$  关于  $h_{ij}$  的其他高阶导数为  $F^{(n)}_{h_{ij}}(h) = 0, n \geq 5$ . 令  $h'_{ij}$  表示第  $t$  次迭代更新后的  $h_{ij}$  值, 则  $F_{h_{ij}}(h)$  在  $h'_{ij}$  处的泰勒展开式为

$$\begin{aligned} F_{h_{ij}}(h) &= F_{h_{ij}}(h'_{ij}) + F'_{h_{ij}}(h'_{ij})(h - h'_{ij}) + \frac{1}{2}F''_{h_{ij}}(h'_{ij})(h - h'_{ij})^2 + \frac{1}{6}F^{(3)}_{h_{ij}}(h'_{ij})(h - h'_{ij})^3 + \\ &\quad \frac{1}{24}F^{(4)}_{h_{ij}}(h'_{ij})(h - h'_{ij})^4 \end{aligned} \quad (\text{A6})$$

定义函数

$$\begin{aligned} G_{h_{ij}}(h, h'_{ij}) &= F_{h_{ij}}(h'_{ij}) + F'_{h_{ij}}(h'_{ij})(h - h'_{ij}) + \frac{1}{2} \left[ \frac{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3}{h'_{ij}} \right] \times \end{aligned}$$

$$(h - h'_{ij})^2 + \frac{1}{6} F_{h_{ij}}^{(3)}(h'_{ij})(h - h'_{ij})^3 + \frac{1}{24} F_{h_{ij}}^{(4)}(h'_{ij})(h - h'_{ij})^4$$

是  $F_{h_{ij}}(h)$  的一个辅助函数。

证明. 当  $h'_{ij} = h$  时, 有  $G_{h_{ij}}(h, h) = F_{h_{ij}}(h)$ , 下面证明: 当  $h'_{ij} \neq h$  时, 有  $G_{h_{ij}}(h, h'_{ij}) \geq F_{h_{ij}}(h)$ . 对比  $F_{h_{ij}}(h)$  和  $G_{h_{ij}}(h, h'_{ij})$ , 可以得到要证明  $G_{h_{ij}}(h, h'_{ij}) \geq F_{h_{ij}}(h)$ , 需证明如下不等式成立:

$$\frac{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3}{h'_{ij}} \geq F_{h_{ij}}''(h) \quad (\text{A7})$$

由于  $h'_{ij} \geq 0, w_{ij} \geq 0$  以及  $\mathbf{X}_{ii} \geq 0$ , 故有

$$\begin{aligned} & \frac{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3}{h'_{ij}} \\ &= 2(\mathbf{H}\mathbf{H}^T)_{ii} + 2 \sum_{k \neq i} (h_{kj})^2 + 4(h_{ij})^2 + (\mathbf{W}^T\mathbf{W})_{jj} + \lambda \\ &\geq -2\mathbf{X}_{ii} + 2((\mathbf{H}\mathbf{H}^T)_{ii} + h_{ij}^2 + \sum_k h_{kj}^2) + (\mathbf{W}^T\mathbf{W})_{jj} + \lambda \\ &= F_{h_{ij}}''(h). \end{aligned}$$

所以式(A7)成立, 根据正文的定义 1 可知  $G_{h_{ij}}(h, h'_{ij})$  是  $F_{h_{ij}}(h)$  的辅助函数. 下面证明  $F_{h_{ij}}(h)$  的迭代求解规则符合正文的引理 1. 根据正文的引理 1 可知只要能够获得  $G_{h_{ij}}(h, h'_{ij})$  的局部极小点, 就可以得到  $F_{h_{ij}}(h)$  的局部极小点, 而  $G_{h_{ij}}(h, h'_{ij})$  的局部极小点可以通过求解方程  $G'_{h_{ij}}(h, h'_{ij}) = 0$  获得. 由于

$$\begin{aligned} G'_{h_{ij}}(h, h'_{ij}) &= F'_{h_{ij}}(h'_{ij}) + \left[ \frac{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3}{h'_{ij}} \right] \times \\ & \quad (h - h'_{ij}) + \frac{1}{2} F_{h_{ij}}^{(3)}(h'_{ij})(h - h'_{ij})^2 + \frac{1}{6} F_{h_{ij}}^{(4)}(h'_{ij})(h - h'_{ij})^3 \end{aligned} \quad (\text{A8})$$

其中  $G'_{h_{ij}}(h, h'_{ij})$  是关于  $h$  的泰勒展开式, 因此为求出方程  $G'_{h_{ij}}(h, h'_{ij}) = 0$  的解, 可以采用牛顿迭代求解公式, 即

$$h_{ij}^{t+1} = h_{ij}^t - \frac{G'_{h_{ij}}(h_{ij}^t, h'_{ij})}{G''_{h_{ij}}(h_{ij}^t, h'_{ij})} \quad (\text{A9})$$

其中,

$$G'_{h_{ij}}(h_{ij}^t, h'_{ij}) = (-2\mathbf{X}\mathbf{H} + 2\mathbf{H}\mathbf{H}^T\mathbf{H} - \mathbf{Y}^T\mathbf{W} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} \quad (\text{A10})$$

$$G''_{h_{ij}}(h_{ij}^t, h'_{ij}) = \frac{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3}{h'_{ij}} \quad (\text{A11})$$

可求得

$$h_{ij}^{t+1} = \frac{(2\mathbf{X}\mathbf{H} + \mathbf{Y}^T\mathbf{W})_{ij} + 4(h_{ij})^3}{(2\mathbf{H}\mathbf{H}^T\mathbf{H} + \mathbf{H}\mathbf{W}^T\mathbf{W} + \lambda\mathbf{H})_{ij} + 4(h_{ij})^3} h_{ij}^t \quad (\text{A12})$$

牛顿迭代求解公式可以保证  $G'_{h_{ij}}(h, h'_{ij})$  收敛, 那么  $G_{h_{ij}}(h, h'_{ij})$  可以获得局部极小点, 根据正文的引理 1,  $F_{h_{ij}}(h)$  也可以获得局部极小点, 即在式(A12)表示的迭代求解规则下更新  $\mathbf{H}$  的所有元素, 可以使目标函数  $F$  不断收敛, 并最终获得局部最小值. 对比式(A1)和式(A12)的迭代更新规则, 事实上两者是等价的, 因为式(A12)的共同项  $4(h_{ij})^3$  消除后

并不影响更新规则的收敛性, 所以通过以上证明可知式(A1)可以作为矩阵  $\mathbf{H}$  的迭代更新规则. 同理, 对于  $\forall w_{pq} \in \mathbf{W}$ , 用函数  $F_{w_{pq}}(w)$  表示  $F$  中跟  $w_{pq}$  相关的部分, 可以计算  $F_{w_{pq}}(w)$  在  $w_{pq}$  的一阶至三阶偏导数分别为  $F'_{w_{pq}}(w) = (-\mathbf{Y}\mathbf{H} + \mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}$ ,  $F''_{w_{pq}}(w) = (\mathbf{H}^T\mathbf{H})_{pq} + \varphi$  以及  $F_{w_{pq}}^{(3)}(w) = 0$ . 令  $w'_{pq}$  表示第  $t$  次迭代更新后的  $w_{pq}$  值, 则  $F_{w_{pq}}(w)$  在  $w'_{pq}$  处的泰勒展开式为

$$F_{w_{pq}}(w, w'_{pq}) = F_{w_{pq}}(w'_{pq}) + F'_{w_{pq}}(w'_{pq})(w - w'_{pq}) + \frac{1}{2} F''_{w_{pq}}(w'_{pq})(w - w'_{pq})^2 \quad (\text{A13})$$

定义函数

$$G_{w_{pq}}(w, w'_{pq}) = F_{w_{pq}}(w'_{pq}) + F'_{w_{pq}}(w'_{pq})(w - w'_{pq}) + \frac{1}{2} \left[ \frac{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}{w'_{pq}} \right] (w - w'_{pq})^2$$

是  $F_{w_{pq}}(w)$  的辅助函数.

证明. 当  $w'_{pq} = w$  时, 有  $G_{w_{pq}}(w, w) = F_{w_{pq}}(w)$ , 下面证明: 当  $w'_{pq} \neq w$  时, 有  $G_{w_{pq}}(w, w'_{pq}) \geq F_{w_{pq}}(w)$ . 对比  $F_{w_{pq}}(w)$  和  $G_{w_{pq}}(w, w'_{pq})$ , 可以得到要证明  $G_{w_{pq}}(w, w'_{pq}) \geq F_{w_{pq}}(w)$ , 需证明如下不等式成立:

$$\frac{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}{w'_{pq}} \geq F''_{w_{pq}}(w) \quad (\text{A14})$$

由于  $w'_{pq} \geq 0$  以及  $h_{pq} \geq 0$ , 故有

$$\begin{aligned} \frac{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}{w'_{pq}} &= (\mathbf{H}\mathbf{H}^T)_{qq} + \sum_{k \neq q} w_{pk} (\mathbf{H}^T\mathbf{H})_{kq} + \varphi \\ &\geq 2(\mathbf{H}\mathbf{H}^T)_{qq} + \varphi \\ &= F''_{w_{pq}}(w). \end{aligned}$$

所以式(A14)成立, 根据正文的定义 1 可知  $G_{w_{pq}}(w, w'_{pq})$  是  $F_{w_{pq}}(w)$  的辅助函数. 下面证明  $F_{w_{pq}}(w)$  的迭代求解规则符合正文的引理 1. 根据正文的引理 1 可知只要能够获得  $G_{w_{pq}}(w, w'_{pq})$  的局部极小点, 就可以得到  $F_{w_{pq}}(w)$  的局部极小点, 而  $G_{w_{pq}}(w, w'_{pq})$  的局部极小点可以通过求解方程  $G'_{w_{pq}}(w, w'_{pq}) = 0$  获得. 由于

$$G'_{w_{pq}}(w, w'_{pq}) = F'_{w_{pq}}(w'_{pq}) + \left[ \frac{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}{w'_{pq}} \right] (w - w'_{pq}) \quad (\text{A15})$$

$G'_{w_{pq}}(w, w'_{pq})$  是关于  $w$  的泰勒展开式, 因此为求出方程  $G'_{w_{pq}}(w, w'_{pq}) = 0$  的解, 可以采用牛顿迭代求解公式, 即

$$w_{pq}^{t+1} = w_{pq}^t - \frac{G'_{w_{pq}}(w_{pq}^t, w'_{pq})}{G''_{w_{pq}}(w_{pq}^t, w'_{pq})} \quad (\text{A16})$$

其中,

$$G'_{w_{pq}}(w_{pq}^t, w'_{pq}) = (-\mathbf{Y}\mathbf{H} + \mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq} \quad (\text{A17})$$

$$G''_{w_{pq}}(w_{pq}^t, w'_{pq}) = \frac{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}}{w'_{pq}} \quad (\text{A18})$$

则有

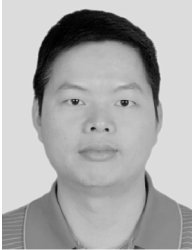
$$w_{pq}^{t+1} = \frac{(\mathbf{Y}\mathbf{H})_{pq}}{(\mathbf{W}\mathbf{H}^T\mathbf{H} + \varphi\mathbf{W})_{pq}} w_{pq}^t \quad (\text{A19})$$

牛顿迭代求解公式可以保证  $G'_{w_{pq}}(w, w'_{pq})$  收敛, 那么  $G_{w_{pq}}(w, w'_{pq})$  可以获得局部极小点. 根据正文的引理 1,  $F_{w_{pq}}(w)$  也可以获得局部极小点, 即在式(A19)表示的迭代



规则下更新  $\mathbf{W}$  的所有元素, 可以使目标函数  $F$  不断收敛, 并最终获得局部最小值. 对比式 (A2) 和式 (A19) 的迭代更新规则, 两者相同, 因此式 (A2) 可以作为矩阵  $\mathbf{W}$  的迭代更新规

则. 综上证明可知应用迭代更新规则式 (A1) 和 (A2) 可以保证  $F$  为非增函数, 并且可以收敛获得最小值,  $\mathbf{H}$  和  $\mathbf{W}$  可以获得局部最优解.



**HE Chao-Bo**, born in 1981, Ph. D. , associate professor. His research interests include data mining and social computing.

**TANG Yong**, born in 1964, Ph. D. , professor, Ph. D. supervisor. His research interests include cloud computing and big data.

**LIU Hai**, born in 1974, Ph. D. , associate professor. His research interests include machine learning and semantic web.

**ZHAO Gan-Sen**, born in 1977, Ph. D. , professor. His research interests include cloud computing and information security.

**CHEN Qi-Mai**, born in 1965, M. S. , professor. His research interests include data warehouse and data mining.

**HUANG Chang-Qin**, born in 1972, Ph. D. , professor. His research interests include service computing and semantic web.

Background

Community mining in complex network has attracted lots of attentions and many methods have been proposed to solve this problem. However, most of exiting methods only consider the usage of node link or attribute information and cannot mine community where members link densely and have nearly the same attributes. In fact, this type of community is very useful for many applications, such as topic community discovery and social recommendation in online social network. For achieving this goal, community mining should consider the usage of node link and attribute information at the same time.

There have some works focusing on mining community by integrating link and attribute information. Some related papers have been presented in some famous academic conferences, such as SIGMOD, KDD and WWW. However, these works have nearly the same problems: cannot integrate together link and attribute information using the same model and need to use other cluster algorithms to obtain final results. Aiming at these problems, we propose a method named LANMF which is based on joint nonnegative matrix factorization (NMF) model and can factorize uniformly nodes link matrix and attributes association matrix. Community membership can be directly obtained based on the community membership factorization matrix. And, the factorization matrix of attributes associated strength can be directly used to describe representative topics of every community. By

using the powerful cluster utility of joint NMF model, relevancies of nodes link structure and attributes in each community can be well guaranteed. Furthermore, by using the optimization algorithm LANMF has the linear time complexity and can be used to mine community in the large-scale complex network. We conduct extensive experiments in some real world complex network datasets and the results show that our method outperforms state-of-the-art methods and is also very effective to mine topic community and overlapped community existing in social network.

This work is supported by the National High-Technology Research and Development Program (863 Program) of China (No. 2013AA01A212), the National Natural Science Foundation of China (Nos. 61370178, 61370229), the Natural Science Foundation of Guangdong Province, China (Nos. S2012030006242, 2015A030310509), the Science and Technology Support Program of Guangdong Province, China (Nos. 2016A030303058, 2016A020210130, 2015A020209178, 2015B010129009, 2014B010103004, 2014B010117007, 2015A030401087, 2015B010110002), the Open Research Fund from Guangzhou Key Lab on Cloud Security and Assessment (No. GZCSKL-1407). Community mining is one of key problems concerned by these projects. The theoretical results of this paper can provide solid foundation for their further research.