基于图流在线非负矩阵分解的社团检测

常振超,陈鸿昶,王 凯,卫红权,黄瑞阳
(国家数字交换系统工程技术研究中心,河南郑州 450001)

摘 要: 针对现有的在线社团检测方法大多仅从增量相关的节点和边出发,难以有效挖掘社团结构的动态变化特性问题,提出了一种基于图流在线非负矩阵分解的社团检测方法.首先将网络中持续到达的图数据按照流式数据进行存储和预处理,然后借鉴梯度下降思想,采用在线非负矩阵分解架构,根据不同时刻达到的图流序列,实时迭代更新社团归属矩阵,并通过有效的学习率和缓存策略设置,保证了图流处理的收敛性和合理性.实验结果表明,相比于已有在线社团检测方法,该方法具备更高的社团检测精度.

关键词: 在线;非负矩阵分解;图流;社团检测
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2017)09-2077-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.09.004

Graph Streams Community Detection via Online Nonnegative Matrix Factorizations

CHANG Zhen-chao, CHEN Hong-chang, WANG Kai, WEI Hong-quan, HUANG Rui-yang
(*National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, Henan* 450002, *China*)

Abstract: While existing online community detection methods mostly only deal with the nodes and edges which from the increment part, which are difficult to effectively detect the dynamic changes in the community structure. Based on this, a new method for the detection of flow graphs based on online non negative matrix factorization (ONMF) is proposed. Firstly, our method put graph data into the cache as continuous streams to deal with. Then, our method iterative updates the existing community belonging matrix real-time using online nonnegative matrix decomposition architecture and by means of the projected gradient descent theory. Lastly, through effective learning rate and cache strategy setting, our method ensures the convergence and rationality of graph stream processing. Experiments on real network data sets show that ONM has a higher community detection quality compared with existing methods.

Key words: online; nonnegative matrix factorization; graph streams; community detection

1 引言

网络涵盖了物理世界的方方面面,通常可用图来进行描述,节点表示物理世界中的实体,边表示实体之间的联系.对网络图中的社团结构进行挖掘分析是理解网络基本性质和组成的基础,一直是各学科领域研究者所共同关注的热点[1].社团是指具备密集交互的节点组合,如物理学中的凝聚与离散物体的汇合、社交网络中具备相同爱好的群体和引文网络中具备相同领域的研究小组等[2].已有大量的社团检测算法被提出,如基于连接的GN算法[3]、边的信息传播分析方法[4]、非负矩阵分解方法[5,6]等,这些方法大多基于静态网络分析角度出发.

近年来,网络技术的迅猛发展极大的便利了人们之间的沟通与交流,也产生了大规模的网络图数据,如Facebook社交网络已经有超过十亿的用户和上百亿的连接关系[7],twitter也有超过四千万的用户和十亿条连接关系[8].这些海量数据通常都具备"图流"特性,图流指的是构成网络中关系图的数据是动态产生的,即随着时间变化,网络数据依流式特征动态到达,如在线社交网络(twitter、Facebook等)用户的增加和减少、用户之间关系的建立和删除等.大数据时代图流数据实时

收稿日期:2016-01-11;修回日期:2016-03-29;责任编辑:梅志强
基金项目:国家自然科学基金创新群体(No. 61521003);国家自然科学基金(No. 61171108);国家973重点基础研究发展计划(No. 2012CB315901, No. 2012CB315905);国家科技支撑计划(No. 2014BAH30B01)

产生,给社团检测技术带来了新的挑战,需要研究内存有限情况下高效的处理方法. 传统的基于全局静态的分析方法不再适用,采用动态分析的在线学习方法是针对此类问题研究的新趋势.

当前针对图流数据的社团检测研究主要有两个方面:第一类方法是设计并行高效的处理架构,如常见基于多处理器的分布式处理架构、基于线性算法扩展的多核并行处理机制等. 如文献[9]设计了基于 Hadoop 聚类算法平台,实现了大规模网络的并行化计算. Wickramaarachchi 等人[10]通过修改序列 Louvain 算法来实现并行社团检测,通过并行化来解决算法初次迭代中的花销. Prat-Pérez 等人[11]提出了利用多核处理器来提高并行化的方式,通过最大化加权社团聚类完成社团花分. Gregori 等人[12]通过并行化 K-Clique 算法,并通过集成学习方式达到了很好的效果. 但这些方法通常需要将原始数据进行分块处理,无法保证原始数据拓扑结构的完整性,且大多仍是针对静态数据集处理,缺乏在线学习机制. 第二类方法集中在增量实时的处理方式上,大多仅从增量相关的节点和边出发,保证了算法效率的同时,也考虑了社团的动态变化. 如 Yun 等人[13]将社团检测分为在线和离线两个阶段,在线阶段完成对所到达的流式数据的社团划分,离线阶段用于存储所有已处理节点的社团划分. Tsourakakis 等人[14]将流图划分看成是平衡图划分和社团检测这两类相关性问题. 文献[15]将图中的节点和边按照流式数据进行更新,根据先后次序逐步老化数据的权重. Lin 等人[16]提出了一种基于生成模型的增量动态社团检测架构 FaceNet,首次将非负矩阵分解架构应用到动态网络社团检测中来,取得了不错的检测效果. 郭进时等人[17]将拓扑势引入到增量相关节点的处理上去,提高了算法的检测精度. 近两年,也有将经典算法进行扩展为增量聚类的研究,如 Pan 等人[18]提出的 OLEM 方法,从期望模块度增加,按照边到边的局部角度进行研究,取得了较好的效果. Duan 等人[19]提出的增量派系图方法,用于解决社团检测问题. 此类方法通常只对增量相关部分出发,缺乏对已有数据变化的综合考虑.

作为一种有效地维数约简算法工具,非负矩阵分解(NMF)在数据挖掘和信息提取领域得到了广泛的应用. NMF 可以看作是一种宽松化的 k-means 聚类[20],与谱聚类和概率潜在情感分析有着内在的关系[21]. 基于 NMF 的社团检测与其他传统方法相比,也起到了较好的识别效果[5,6]. 但现有的 NMF 算法只是针对网络是静态图的情况下的社团检测,其本身运算复杂度高,且需要将整张图读入内存进行处理,针对大规模图的处理情况,基于在线的非负矩阵分解(ONMF)社团检测方法有待进行研究. 作为一种有效地在线的处理机制,ONMF 主要用于解决大规模数据集中全局数据无法获取情况,它从减少内存开销和增加运算效率出发,有效应对了流式数据处理问题,在文本和图像挖掘领域已有了广泛的应用. 最早的 ONMF 方法由 Cao 等人[22]提出,将其用于跟踪随时间演化数据中的潜在变化主题,有效获取了主题随时间的变化情况. 后续研究也将 ONMF 进行扩展到了大规模数据集的挖掘任务上. 如 Wang 等人[23]将 ONMF 用于处理大规模流式 twitter 数据中的文本聚类,克服了传统 NMF 需要整个数据集进行运算的缺陷. Guan 等人[24]将 ONMF 用于大规模图流数据中的图像搜索中,也取得了较好的效果

基于上述分析,针对现有算法无法有效应对网络规模持续膨胀的问题,本文从在线图流处理角度研究出发,提出了一种基于图流在线非负矩阵分解的社团检测架构(Online Nonnegative Matrix Factorization, ONMF),该方法以在线非负矩阵分解为社团逼近方法,按照在线的增量学习方式,对每一组到达的样本集合进行处理,实时更新子空间中分解得到的基矩阵,进而获取动态的社团结构.

2 在线图流社团检测模型

本文采用在线非负矩阵架构对社团进行划分,针对新到达的流图数据应用非负矩阵分解,并更新历史时刻得到的划分.

2.1 相关定义

图流是指由不同的离散时刻 $t_1,t_2,\cdots,t_m$ 所构成的网络快照序列 $G=\{G_1,G_2,\cdots,G_m\}$,m 为序列的个数. 在 t_i 时刻增量网络可由 $G_i=(V_i,E_i)$ 表示,其中 V_i 表示 t_i 时刻网络中的节点集合,$V_i=\{v_{i1},v_{i2},\cdots,v_{in}\}$ 为 t_i 时刻内达到的 n 个节点,E_i 表示 t_i 时刻到达节点构成网络的边集合,$E_i=\{e_{i1},e_{i2},\cdots,e_{im}\}$ 表示 t_i 时刻网络中的 m 条边. 某个流图可以指某个时间间隔内的邮件、电话和社交网络交互关系构成的图. 在线网络社团划分由离散时刻的社团子集所构成 $C=\{C_1,C_2,\cdots,C_m\}$,C_i 为 t_i 时刻获取的网络社团划分集合.

定义 1 t_i 时刻邻接矩阵 $\boldsymbol{X}_i\in\mathbf{R}^{n\times n}$,$n$ 为节点总数,矩阵元素的值为其中两个节点之间的链接关系,即当节点 v_i 和节点 v_j 之间有链接时,$\boldsymbol{X}_i(i,j)=1$;反之,$\boldsymbol{X}_i(i,j)=0$. 通常情况下,社会网络中 $\boldsymbol{G}$ 为稀疏矩阵.

定义 2 t_i 时刻归属矩阵 $\boldsymbol{H}_i\in\mathbf{R}^{K\times n}$,$\boldsymbol{H}_i$ 的第 j^{th} 列表示节点 v_j 在 K 个社团上的归属程度,刻画了原始信息的结构与特征.

定义 3 t_i 时刻社区特征矩阵 $\boldsymbol{W}_i\in\mathbf{R}^{n\times K}$,$\boldsymbol{W}_i$ 的表示网络降维后的社团基特征分布.

在给定网络信息的基本定义之后,在线社团检测问题为:在 t_i 时刻到达的图流数据构成的连接关系 $\boldsymbol{A}_i$

$\in \mathbf{R}^{n\times n}$,以及 t_{i-1} 时刻获取的潜在因子矩阵 $\boldsymbol{H}_{i-1}$ 和 $\boldsymbol{W}_{i-1}$,更新并获取新的归属矩阵 $\boldsymbol{H}_i$ 和特征矩阵 $\boldsymbol{W}_i$,进而能够更为准确的反映所有到达数据的总体分布特性. 本文在线图流社团检测如图 1 所示.

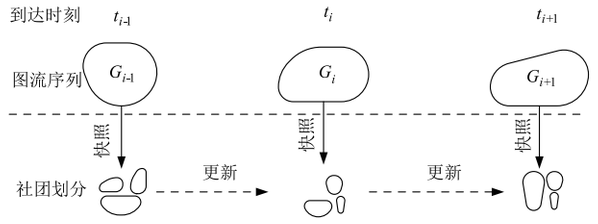


图1　在线图流社团检测示意图

2.2　基于 NMF 的社团检测

由于网络节点间的链接是非负的,即边权重都是非负的,因此,非常适合采用非负矩阵分解进行社团检测. 基于非负矩阵进行社团检测的基本定义如下,假设拥有 n 个节点的网络 $G(V,E)$ 的邻接矩阵为 $\boldsymbol{X}\in\mathbf{R}^{n\times n}$,则 NMF 定义为:通过寻找最大近似原始网络数据 X 的 2 个低秩因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 来实现社区发现,分解后得到的基向量矩阵 $\boldsymbol{W}$ 表示网络降维后的聚类社区特征,而归属矩阵 $\boldsymbol{H}$ 则表示相应节点在社区中的隶属程度,矩阵求解过程采用欧几里德距离最小化方式,优化的目标函数 $O^l(E)$ 为

$$\begin{aligned}&\min\ O^l(E)=\min_{W,H}\|\boldsymbol{X}-\boldsymbol{WH}\|_{\mathrm{F}}^2\\&\text{s.t.}\quad \boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0\end{aligned}\tag{1}$$

其中 $\|\cdot\|_{\mathrm{F}}$ 为 Frobenius 范数(简称 F 范数),用来度量目标函数的逼近程度;$\boldsymbol{W}\in\mathbf{R}^{n\times r}$ 和 $\boldsymbol{H}\in\mathbf{R}^{r\times n}$ 分别是分解之后得到的关于模式节点的基矩阵和归属矩阵,n 表示网络中的节点个数,r 表示相关模式节点子空间的聚类个数,即网络 G 中存在的社区个数.

2.3　采用 ONMF 的社团检测方法

NMF 目的是用一组低维非负基向量的线性组合来描述高维的输入数据. 而 ONMF 的思想是动态地改变非负基向量,以更好地描述不断增加的网络数据中所包含的社团结构. 由于历史时刻获取的基向量能够用于描述已经到达的网络数据,因此,当新的流数据到达时,无需重新对所有的数据进行处理,仅需要根据新时刻到达的数据进行更新已有基向量,进而以增量在线的方式描述所有网络数据.

基于流图的社团检测,需要处理的对象是持续膨胀的网络数据,即按照时间序列到达的流式数据. 假设每个流图序列所占用的空间为 $P\in\mathbf{R}^n$,即给定的 m 个时刻达到流图序列 $\{V_1,\cdots,V_m\}\in\mathbf{R}^n$,NMF 的目的是找到降维后的子空间 $Q\in\mathbf{R}^n$. 子空间包含 r 个基子向量 $\{w_1,\cdots,w_r\}$ 构成,r 是经矩阵分解后的所用于描述得聚类中心的基矩阵 $\boldsymbol{W}$ 维数,即社团数目,分解的得到的另一个矩阵 $\boldsymbol{H}$ 为概率矩阵,描述了节点对于社团归属程度. 在线社团检测每次处理序列中的一个流图,根据当前时刻新到达的流图数据,增量更新基矩阵 $\boldsymbol{W}$ 和社团归属矩阵 $\boldsymbol{H}$. 其求解过程可用下式表示:

$$\Psi(\boldsymbol{W},\boldsymbol{H})=\|\boldsymbol{X}-\boldsymbol{WH}\|_{\mathrm{F}}^2=\sum_{i=1}^{m}\|\boldsymbol{X}_i-\boldsymbol{W}_i\boldsymbol{H}_i\|_{\mathrm{F}}^2\tag{2}$$

假设第 t_i 时刻网络中到达的数据流构成的图用邻接矩阵 $\boldsymbol{X}_i$ 表示,与式(1)类似,由于目标是非凸的,为保证优化目标有解,ONMF 对 $\boldsymbol{W}_i$ 和 $\boldsymbol{H}_i$ 分别进行优化求解,即固定其中一个变量,完成对另一个变量的求解. 在本文的在线算法设计中,对 t_i 时刻的数据进行处理时,首先根据到达的流数据构造邻居矩阵 $\boldsymbol{X}_i$,然后将之前时刻获取的基矩阵 $\boldsymbol{W}_{i-1}$ 进行固定,通过最小化逼近误差,以获取该时刻的归属矩阵 $\boldsymbol{H}_i$,其优化目标如下式所示:

$$\min_{H\in\mathbf{R}^r}\|\boldsymbol{X}_i-\boldsymbol{W}_{i-1}\boldsymbol{H}_i\|_{\mathrm{F}}^2\tag{3}$$

相应地,获取 $\boldsymbol{H}_i$ 之后,$\boldsymbol{W}_i$ 可以通过式(4)获取:

$$\boldsymbol{W}_i=\arg\min_{W\in\mathbf{R}^{m\times r}}E_X\|\boldsymbol{X}_i-\boldsymbol{W}_i\boldsymbol{H}_i\|_{\mathrm{F}}^2\tag{4}$$

在对优化目标式(4)的求解过程,借鉴经典的随机梯度下降法[25]思想,计算目标函数的梯度值,并根据前一次迭代获取的矩阵因子,来完成本次迭代矩阵因子的估计. 以 t_i 时刻梯度下降法求解 $\boldsymbol{W}_i$ 过程为例,已知之前时刻得到的矩阵 $\boldsymbol{H}_{i-1}$ 和 $\boldsymbol{W}_{i-1}$,并且获取了流图数据的邻接矩阵 $\boldsymbol{X}_i$,将求解 $\boldsymbol{W}_i$ 初次迭代的 $\boldsymbol{W}_{(i)}^{(0)}$ 初始化为 $\boldsymbol{W}_{i-1}$,已知 $\boldsymbol{H}$ 定义优化 $\boldsymbol{W}$ 的目标函数如下:

$$F(\boldsymbol{W},\boldsymbol{X})=\|\boldsymbol{X}-\boldsymbol{WH}\|_{\mathrm{F}}^2\tag{5}$$

由文献[25]可知,基于定义式(5),借鉴文献[24]中鲁棒的随机近似方法(RSA)来优化式(4),通过使用灵活的学习率选择技巧和平均化准则,来保证算法的收敛性保持在 $O(\frac{1}{\sqrt{k}})$. RSA 随机利用了 N 个已到达的图流序列构造的样本空间 $\{X_1,\cdots,X_N\}$,来对当前的 $\boldsymbol{W}_i$ 进行递归更新,如下式子所示:

$$\boldsymbol{W}_i^{(k+1)}=\phi(\boldsymbol{W}_i^{(k)}-\alpha^{(k)}\nabla_W F_i(\boldsymbol{W}_i^{(k)},\boldsymbol{X}_k))\tag{6}$$

其中 ϕ 表示梯度映射的非负约束空间,$\nabla_W F_i(\boldsymbol{H}_{(i)}^{(k)},\boldsymbol{X}^{(k)})$ 为对式(4)目标函数中的 $\boldsymbol{W}$ 进行梯度求解,K 的取值为 $1\sim N$,$\alpha^{(k)}$ 是步长准则,由 Armijo 准则[26]所约束,取值范围为区间(0,1),其定义为:

$$\alpha^{(k)}=\theta_i(D_W/M_*)\sqrt{k}\tag{7}$$

其中 $D_W=\max\|\boldsymbol{W}-\boldsymbol{W}_1\|_{\mathrm{F}}$,$\theta_i$ 是第 i 次数据到达的学习率缩放比例值,决定了算法求解的收敛速率,将其取值一般较小,$M_*=\sup_{W\in C}E_X^{1/2}(\|\nabla(\boldsymbol{W},\boldsymbol{H})\|_{\mathrm{F}}^2)$.

在针对流式进行连续处理的步骤中,将 $\boldsymbol{W}_0$ 初始化为随机矩阵,$\boldsymbol{W}_{i-1}$ 为 t_i 的前一时刻获取的基矩阵,针对 m 个时刻到达的图流序列 $\{G_1,\cdots,G_m\}$,能够根据上一

个时刻的基矩阵求解系数矩阵 $\boldsymbol{H}$ 序列$\{\boldsymbol{H}_1,\cdots,\boldsymbol{H}_m\}$,进而实时更新基矩阵 $\boldsymbol{W}$ 序列$\{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_m\}$.

优化问题(5)中 $\boldsymbol{W}_i$ 求解的停机准则设定如下:

$$\frac{\|\boldsymbol{W}_i^k-\boldsymbol{W}_i^{k-1}\|_{\mathrm{F}}}{\|\boldsymbol{W}_i^{k-1}\|}\leqslant\tau \tag{8}$$

其中 τ 是相邻两次迭代获取的 $\boldsymbol{W}_i$ 相差值,其取值一般较小,通常设定为 10^{-5}.

在针对流式数据处理上,为了有效存储实时到达的流图数据,本文引入缓存模块 B,用于存储当前时刻到达的采样和之前获取的系数矩阵序列. 这样算法的空间复杂度仅于存储在缓存模块 B 中存储空间有关,将原始不断增长的流式数据处理,看作每次处理缓存模块 B 中的数据. 本文所提基于 ONMF 的图流社团检测算法流程如算法 1 所示.

算法 1 在线非负矩阵分解的图流社团检测

输入:时刻 t_1 到 t_m 的流图序列 $G_0,G_1,\cdots,G_m$,采样时刻 T,基特征个数 r.

输出:时刻 $t_i(1\leqslant i\leqslant m)$ 的基特征 $\boldsymbol{W}_i\in\mathbf{R}^{n\times K}$,时刻 t_i 社团归属矩阵 $\boldsymbol{H}_i\in\mathbf{R}^{r\times n}$.

1. 参照经典的非负矩阵分解设定[25],随机初始化 $\boldsymbol{H}_0$ 和 $\boldsymbol{W}_0$ 为非负值,区间为(0,1)之间,并初始化 $B\leftarrow\phi$

For $t_i=t_1:t_m$ do

2. 根据采样时刻到达的流数据,构造 $\boldsymbol{X}_i$

3. 由式(3),更新并获取 $\boldsymbol{H}_i$

4. 将 $\boldsymbol{X}_i$ 存储到缓存 B 中

5. 初始化 $\alpha^{(1)}$ 和 τ 值大小,初始化 $\boldsymbol{W}_i^{(1)}\leftarrow\boldsymbol{W}_{i-1}$

Repeat

6. 更新 $\alpha^{(k)}$,$k\leftarrow 1$

7. 从 B 中获取 $\boldsymbol{X}_i$

8. 根据式(6),更新 $\boldsymbol{W}_i$

9. $k=k+1$

Until 满足停机准则式(8)

10. 返回 $\boldsymbol{W}_i$ 和 $\boldsymbol{H}_i$

End for

2.4 参数设置

本文算法设计中,需要设定的模型参数有两个,梯度下降的学习率的缩放比例 θ_i 和缓存参数 B.

在式(7)中,θ_i 是第 i 次数据到达的学习率缩放比例值,决定了算法求解的收敛速率,参考文献[24]中设置,将其取值设定在(0,0.1)之间,其设定如下:

$$\theta_i=0.1\cos((t-1)\pi/2T) \tag{9}$$

缓存参数 B. 流式数据处理中,由于数据的无限增长特性,无法将数据完全存储到内存中进行处理,因此需要设定一个缓存,用于存储需要处理的流图数据,并需要实时老化已经处理完毕的流数据. 在 $\boldsymbol{W}$ 的更新计算中,式(6)的迭代更新中,需要存储最近邻的流图序列,假设需要存储最近 l 个时间间隔内到达的样本和已经获取的历史时刻参数,由算法 1 的处理流程可知,算法总的空间复杂度为 $O(nl+rl+nr)$. 针对缓存空间设置中,l 的取值很关键,其取决于两个方面:(1)取值不能太大,太大了增加了空间运算的复杂度;(2)取值要尽量大,能够满足算法 1 中 RSA 的样本空间用于平均化 $\boldsymbol{W}$ 的运算需求,即满足迭代过程有解. 参照文献[24]设置,其取值设定如下:

$$l=\min\{[n/10],20\} \tag{10}$$

其中$[x]$表示取小于 x 的最小整数.

2.5 复杂度分析

本文复杂度分为两个部分进行分析:运算复杂度和空间复杂度. (1)运算复杂度:其主要体现在单次图流的处理上,由于借鉴了随机梯度下降非负矩阵算法求解过程,由文献[24,25]可知,该部分运算复杂度主要集中在基矩阵更新过程,该过程与梯度求解和迭代次数相关,其中式(5)中梯度求解的算法复杂度为 $O(nr)$,其中 n 是单个图流中的节点个数,r 是基矩阵维数,即求得的社团数目,设 K 是更新基矩阵过程的迭代次数,则算法针对单个图流运算复杂度大致为 $O(Knr)$. (2)空间复杂度:空间复杂度在 2.4 节缓存 B 的设置中已经进行了简要分析,其值为 $O(nl+rl+nr)$. 综上分析可知,算法的复杂度与所处理的时刻 T 无关,保持恒定,能够有效应对数据不断增长的处理需求.

3 实验

为验证本文所提算法针对社团结构的检测精度和处理有效性,本文在真实网络数据集上进行了相关的实验,并对实验结果进行了分析.

3.1 实验数据

对已有文献中所广泛应用的网络仿真数据集进行分析,选择了 3 种最常见的大规模数据集进行了仿真实验,数据集来源于斯坦福大学的大规模网络数据集,包括 Amazon 购物网络[27]、DBLP 电子文献网络[28]和 Twitter 社交网络[28]. 这些网络数据集均为大规模的网络数据集,借鉴文献[27]中的实验数据的静态图构建方式,对数据进行提炼,以构造有效的数据集,3 种数据集的详细介绍如下.

Amazon 购物网络:节点为网站上所销售的产品,如果两个品时同时被顾客所购买,则认为这两个产品之间存在一条边. 产品社团归属为已有的商品分类情况. 移除小于 3 个节点的无效链接,该数据集共包含 5000 个社团,共 334863 个节点和 925872 条边.

DBLP 电子文献网络:DBLP 网络为科学文献网络数据库,本文中采用的数据集为计算机科学方面的文献. 节点描述为研究者,边为两个研究者同时出现在同一个电子文献中. 在某个期刊或者会议上发表文献的

全部作者定义为同一个社团全局依据. 同样采集了5000 个社团,节点数为 317080,边数为 1049866.

Twitter 社交网络:Twitter 为典型的社交网络,以 Twitter 用户为节点,用户之间的互动操作为边,以朋友圈作为全局社团依据. 再对数据进行提炼之后,采集了4869 个社团,包含了 81306 个节点和 1768149 条边.

参照文献[18]中的 3 种数据预处理机制,对原始数据集合进行预处理. 由于数据集合均为百万级别的边,因此,本文中按照处理边的增量,将数据集分为多个不同的图序列,具体的边增量设定为 1 万条,这样为每个数据集搭建了 100 个左右的流图序列,进而构造了实验过程所处理的流式动态图序列.

3.2　对比算法

为全面分析比较所提算法基于 ONMF 流图社团检测的性能,实验中挑选了三类具备代表性的对比算法:(1)NMF[5],仅针对静态图的进行社团发现,该方法是一种无监督聚类的方法,由于本文是采用在线的 NMF 方法,故选择仅针对静态快照的 NMF 方法作为对比算法. (2) FaceNet[16],该方法基于生成模型进行演化分析,首次将矩阵分解应用到此类问题中来,其针对历史时刻和当前时刻分别进行联合估计获取社团划分,采用平衡因子将两者划分进行统一,取得了较好的效果,是常用于对比增量社团检测算法中的经典算法. (3) OLEM[18],是一种比较新的在线社团检测方法,具备线性复杂度,适用于大规模的数据处理需求.

3.3　评价指标

本文中数据集采用真实的网络数据集,且各个数据集的构造过程中,节点的社团归属情况是已知的,为已知全局背景情况. 为验证算法那的准确性和有效性,本文中选择常见的成对 F 测度[18]和运行时间作为本文的算法评价指标.

本文中成对 F 测度为最常见的算法性能描述指标,其从两个方面进行综合对算法指标描述,聚类准确度和召回率. 其物理意义为:准确性为检测到正确的节点在已知社团中的比例,召回率为检测到的正确划分的节点在整个检测数据集中的比例. 针对原始数据分析,社团中广泛存在重叠现象,应考虑为重叠社团下正确性描述. 参考文献[18]设定,本文中 H 代表真实情况中至少处于一个社团别中的正确划分的节点对组合,G 代表算法检测出同一个社团中至少出现一次的正确划分节点对组合. 则算法的聚类成对准确度 P 和召回率 R 如下式所示:

$$P = \frac{|H \cap G|}{H} \tag{11}$$

$$R = \frac{|H \cup G|}{H} \tag{12}$$

基于上述表达式,成对测度 F 的定义如下所示:

$$F = \frac{2 \times P \times R}{P + R} \tag{13}$$

由定义可知,成对测度 F 值越高,表明社团检测算法多得到的社团质量越高.

同时,为了验证算法的运行效率,对算法在相同数据集上的运行时间也进行了仿真分析.

3.4　实验结果

在本文的试验中,由于数据没有明显的时间间隔,采用在线方式进行流式处理需要对数据进行时间刻度上的分割,具体数据集和预处理方式见 3.1 节. 3 种数据集上的实验如下面分析所示.

3.4.1　Amazon 购物网络

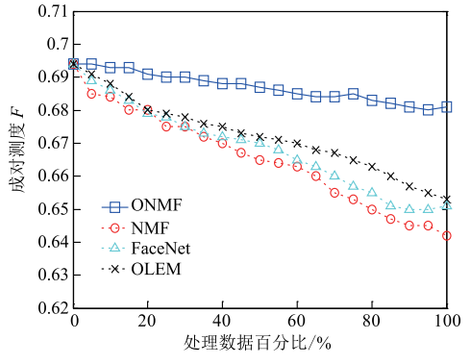
由图 2(a)可知,在对 Amazon 购物网络进行仿真时,基于 ONMF 的社团检测方法,相比于其它 3 种方法,均取得了较高的 F 值. 分析可知,NMF 方法仅对单个静态图进行社团检测的方法,由于缺少对前一时刻信息的综合,F 值较为不稳定,随着数据量的增加,其值最低,效果较差. 随着网络规模的急速增加,网络的社团结构愈发不明显,仅仅依靠对单个静态图的分析,难以有效反映社团结构的变化,基于增量的算法 FaceNet 和 OEM,同样仅对增量相关节点进行划分,难以有效更新整个社团划分结果. 由图 2(b)可知,本文所提的方法在运行时间上保持恒定,因为每次仅仅需要处理单个流图,且在 NMF 迭代求解过程中以上一次结果为基础,通过设定学习率,使得算法较快的收敛,高于其他两种基于矩阵分解的方法 FaceNet 和 NMF. OLEM 仅从增量的节点模块度计算出发,其运行时间也基本保持恒定,但随着网络的不断增加,模块度的计算需要考虑网络社团的规模,因此时间也增加. 总之,ONMF 运行效率与到单次处理图流相关,能够有效应对大规模流式数据的处理场景.

3.4.2　DBLP 电子文献网络

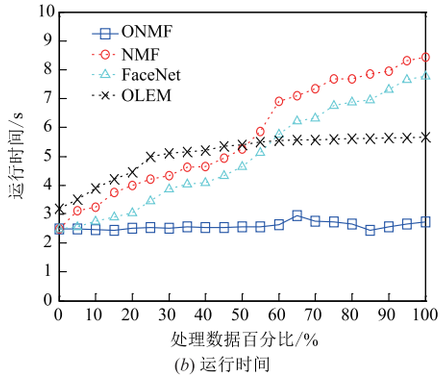
由图 3(a)可知,3 种算法在 DBLP 网络实验结果基本与 Amazon 购物网络相比基本一致,这也验证了基于 ONMF 的社团检测算法能够适用于多种网络结果,在 DBLP 网络的社团检测精度上,所对比 4 种算法中,除 NMF 算法之外,其余 3 种算法检测精度变化很小,说明了 DBLP 而言,随着网络的动态变化,社团结构在随着时间的演进基本变化不大. 由图 3(b)可知,在算法的运算效率上,随着数据量的增加,需要考虑的原始社团划分,因此,除本文 ONMF 算法之外,其余 3 种算法的社团检测时间都呈现出增加趋势.

3.4.3　Twitter 社交网络

由图 4(a)可知,4 种算法在 Twitter 社交网络上的社团检测精度较低,且除 ONMF 算法外,其余 3 种算法

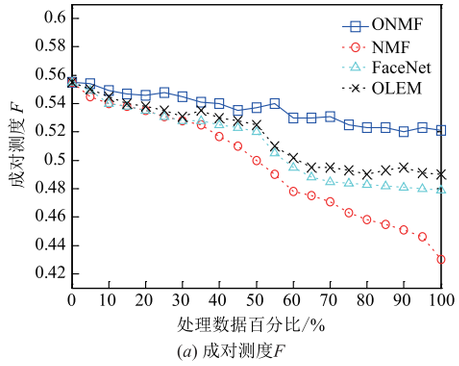


(*a*) 成对测度*F*

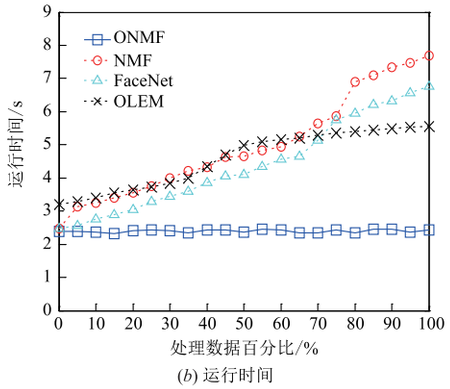


(*b*) 运行时间

图2 Amazon购物网络算法仿真结果



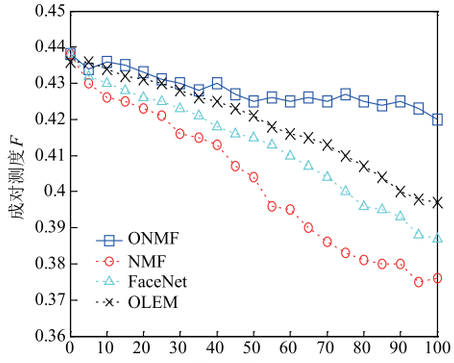
(*a*) 成对测度*F*



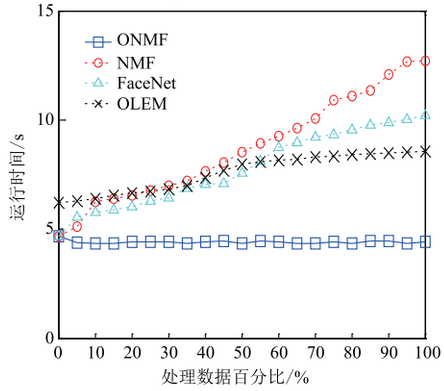
(*b*) 运行时间

图3 DBLP文献网络算法仿真结果

检测精度随着数据量的增加,其精度下降较快. 分析可知,Twitter 为典型的社交网络,其社团结构中度的异质分布情况更为明显,且随着时间的变化,社交网络的用户之间的关系增加和移除不断变化,网络演化交为剧烈. 因此,社团检测的精度普遍较低. 这也验证了本文算法在不同的数据集上的检测精度更具有鲁棒性. 在算法的运行时间上,由图 4(*b*)可知,由于社交网络中节点的度普遍较大,因此,算法运行时间普遍较长,而本文算法仍能保持在一个恒定的运算时间,这也验证了算法的有效性.



(*a*) 成对测度*F*



(*b*) 运行时间

图4 Twitter文献网络算法仿真结果

综合上述 3 种数据集上的实验结果可知,本文算法能够有效适用于购物网络、引文网络和社交网络等不同的真实网络场景. 从算法的检测有效性来说,算法的识别性能均高于其他几类算法,说明了本文所提算法的有效性,能够从持续的流图序列中检测并更新所获取到的社团归属矩阵. 从时间效率来说,本文算法与运算速度较快的增量检测算法相比,算法的处理效率相当,且随着数据集合规模增大,算法处理时间趋于稳定,能够以较为平稳的处理时间来应对流图序列的持续到达情况.

4 结束语

本文从大规模在线图流序列的有效处理角度出

发,提出了一种基于在线非负矩阵分解的社团检测方法.该方法借鉴了梯度下降的思想,利用历史时刻的社团划分和当前所到达的流序列进行实时更新社团检测,为在线社团检测提供了新的研究思路和运算架构,更利于分析网络的动态特性.动态变化的社会网络,为社团检测提供丰富异构的信息源,如何有效利用海量动态的媒体数据信息,将会是下一步工作的研究重点.

参考文献

[1] Fortunato S. Community detection in graphs[J]. Physics Reports,2009,486(3-5):75 – 174.

[2] 潘磊,金杰,等.社会网络中基于局部信息的边社区挖掘[J].电子学报,2012,40(11):2255 – 2263.
PAN Lei, JIN Jie, et al. Detecting link communities based on local information in social networks[J]. Acta Electronica Sinica,2012,40(11):2255 – 2263. (in Chinese)

[3] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2001, 99(12):7821 – 7826.

[4] 张健沛,邓琨,等.基于边标签传播的复杂网络社区识别方法[J].电子学报,2015,43(6):1113 – 1118.
ZHANG Jian-pei, DENG Kun, et al. Community detection in complex networks based on link label propagation[J]. Acta Electronica Sinica, 2015, 43(6): 1113 – 1118. (in Chinese)

[5] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755):788 – 791.

[6] 常振超,陈鸿昶,等.基于联合矩阵分解的节点多属性网络社团检测[J].物理学报,2015,64(21):456 – 465.
Chang Zhen-Chao, Chen Hong-Chang, et al. Community detection based on joint matrix factorization in networks with node attributes[J]. Acta Physica Sinica, 2015, 64(21):456 – 465. (in Chinese)

[7] 皓慧. Facebook CTO 详谈十年愿景:网络连接、AI 和 VR[EB/OL]. http://tech.163.com/16/0817/07/BULDRELV00097U7R.html,2016-08-17.

[8] H Kwak, C Lee, H Park, S Moon. What is twitter, a social network or a news media? [A]. Proceedings of the 19th International Conference on World Wide Web[C]. New York:ACM,2010. 591 – 600.

[9] Varamesh A, Akbari M K, Fereiduni M, et al. Distributed clique percolation based community detection on social networks using MapReduce[A]. 2013 5th Conference on IEEE/ACM Information and Knowledge Technology (IKT)[C]. Shiraz, Iran:IEEE,2013. 478 – 483.

[10] Wickramaarachchi C, Frincu M, Small P, et al. Fast parallel algorithm for unfolding of communities in large graphs[A]. 2014 IEEE High Performance Extreme Computing Conference (HPEC)[C]. Waltham, MA USA:IEEE, 2014. 1 – 6.

[11] Prat-Pérez A, Dominguez-Sal D, Larriba-Pey J L. High quality, scalable and parallel community detection for large real graphs[A]. International Conference on World Wide Web[C]. Seoul, Korea:WWW,2014. 225 – 236.

[12] Gregori E, Lenzini L, Mainardi S. Parallel K-clique community detection on large-scale networks[J]. IEEE Transactions on Parallel & Distributed Systems, 2013, 24(8): 1651 – 1660.

[13] Yun S Y, Lelarge M, Proutiere A. Streaming, memory limited algorithms for community detection[J]. Advances in Neural Information Processing Systems, 2014, 2014(4): 3167 – 3175.

[14] Tsourakakis C, Gkantsidis C, Radunovic B, et al. FENNEL: Streaming graph partitioning for massive scale graphs[A]. Proceedings Of The 7th ACM International Conference On Web Search And Data Mining[C]. New York, USA:ACM,2014. 333 – 342.

[15] Yuan M, Wu K L, Jacques-Silva G, et al. Efficient processing of streaming graphs for evolution-aware clustering[A]. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management[C]. Burlingame, USA:ACM,2013. 319 – 328.

[16] Lin Y R, Chi Y, Zhu S, et al. Analyzing communities and their evolutions in dynamic social networks[J]. Acm Transactions on Knowledge Discovery from Data, 2009, 3(2):307 – 308.

[17] 郭进时,汤红波,王晓雷.基于社会网络增量的动态社区组织探测[J].电子与信息学报,2013,35(9):2240 – 2246.
Guo J, Tang H B, et al. A dynamic community structure detection scheme based on social network incremental[J]. Journal of Electronics & Information Technology, 2013,35(9):2240 – 2246. (in Chinese)

[18] Pan G, Zhang W, Wu Z, et al. Online community detection for large complex networks. [A]. International Joint Conference on Artificial Intelligence[C]. Beijing, China:[U]IJCAI,2013. 1903 – 1909.

[19] Duan D, Li Y, Li R, et al. Incremental K-clique clustering in dynamic social networks[J]. Artificial Intelligence Review, 2012, 38(2):129 – 147.

[20] Chris D, Tao L, Jordan M I. Convex and semi-nonnegative matrix factorizations. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(1):45 – 55.

[21] Gaussier E, Goutte C. Relation between PLSA and NFM and implications[A]. Proceedings of the 28th Annual In-

ternational ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Salvador, Brazil: ACM, 2005. 601 – 602.

[22] Cao B, Shen D, Sun J T, et al. Detect and track latent factors with online nonnegative matrix factorization[A]. International Joint Conference on Artificial Intelligence[C]. Hyderabad, India: IJCAI, 2007. 2689 – 2694.

[23] Wang F, Li P, König A C. Efficient document clustering via online nonnegative matrix factorizations[A]. Proceedings of the 2011 SIAM International Conference on Data Mining[C]. Mesa Arizon, USA: SIAM, 2011. 908 – 919.

[24] Guan N, Tao D, Luo Z, et al. Online nonnegative matrix factorization with robust stochastic approximation[J]. IEEE Transactions on Neural Networks & Learning Systems, 2012, 23(7): 1087 – 1099.

[25] Lin C J. Projected gradient methods for nonnegative matrix factorization[J]. Neural Computation, 2007, 19(10): 2756 – 2779.

[26] Cottle R. Nonlinear programming[J]. John Wiley & Sons, 1979, 4(1): 67 – 110.

[27] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth[J]. Knowledge & Information Systems, 2012, 42(1): 745 – 754.

[28] Mcauley J, Leskovec J. Learning to discover social circles in ego networks[J]. Advances in Neural Information Processing Systems, 2012, 2012: 539 – 547.

作者简介

常振超 男,1987 年生于河北邯郸. 国家数字交换系统工程技术研究中心博士生. 研究方向为网络分析.
E-mail: changzc2012@126.com.cn

陈鸿昶 男,1964 年生于河南郑州. 国家数字交换系统工程技术研究中心教授,博士生导师,研究方向为网络分析.

王 凯 男,1980 年生于河南许昌,国家数字交换系统工程技术研究中心副研究员,研究方向为无网络分析.

卫红权 男,1971 年生于河南郑州,国家数字交换系统工程技术研究中心副研究员,研究方向为无网络分析.

黄瑞阳 男,1986 年生于福建漳州,国家数字交换系统工程技术研究中心讲师,研究方向为无网络分析.