中国科学: 信息科学 2016 年 第 46 卷 第 6 期: 714–728
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
论 文

基于分布式非负矩阵分解的大规模主题社区挖掘

贺超波[①②], 汤庸[②*], 杨阿祧[③②], 赵淦森[②], 刘海[②], 黄昌勤[②]

① 仲恺农业工程学院信息科学与技术学院, 广州 510225
② 华南师范大学计算机学院, 广州 510631
③ 贵州师范大学数学与计算机科学学院, 贵阳 550001
* 通信作者. E-mail: ytang4@qq.com

收稿日期: 2015–11–18; 接受日期: 2016–01–22; 网络出版日期: 2016–05–30
国家自然科学基金 (批准号: 61370178, 61370229)、国家高技术研究发展计划 (批准号: 2013AA01A212)、国家科技支撑计划项目 (批准号: 2012BAH27F05, 2014BAH28F02)、广东省自然科学基金 (批准号: S2012030006242, 2015A030310509)、广东省科技计划项目 (批准号: 2015A020209178)、广东省高性能计算重点实验室开放课题 (批准号: TH1527) 和广州市云计算安全与测评技术重点实验室开放基金 (批准号: GZCSKL-1407) 资助项目

摘要 复杂网络的主题社区挖掘具有重要的应用价值, 但现有方法可扩展性差, 无法高效挖掘大规模复杂网络的主题社区. 针对该问题, 提出一种基于分布式非负矩阵分解的主题社区挖掘方法: TCMDNMF (topic community mining based on distributed nonnegative matrix factorization), 该方法基于非负矩阵联合分解模型, 可以有效统一集成节点链接和内容信息挖掘主题社区. 通过采用梯度下降方法对主题社区挖掘模型进行了优化求解, 并引入 L_1 范数作为稀疏性正则项以及基于 MapReduce 分布式计算框架提高了关键算法的计算效率. 实验结果表明, TCMDNMF 不仅可以有效挖掘主题社区, 而且具有高度可扩展性, 可以有效解决大规模复杂网络主题社区挖掘带来的大数据量计算问题.

关键词 复杂网络 主题社区挖掘 非负矩阵分解 MapReduce 梯度下降

1 引言

复杂网络中的主题社区 (topic community) 挖掘不同于传统的只基于节点链接信息的社区挖掘, 它综合考虑了节点的链接以及内容信息, 从而可以获得成员节点链接紧密且主题特征相似的社区. 主题社区的挖掘结果在现实中更具有应用价值, 如对科技文献合著关系网络的研究方向主题社区挖掘可以发现各研究领域具有代表性的研究团队, 对在线社交网络的兴趣主题社区挖掘可以为电子商务运营商进行精准社会化营销提供有力支持. 目前主题社区挖掘已成为复杂网络社区挖掘问题新的研究热点, 并提出了不少具有代表性的方法, 例如, Zhao 等 [1] 提出一种基于社交对象聚类和链接分析的主题社区发现方法, 该方法首先采用 K-means 算法对节点所关联的社交对象, 如博客、评论等, 进行文本内容聚类并学习节点的主题标签, 然后再通过基于链接信息的模块度最大化社区划分方法挖掘主题社区; Yin 等 [2] 提出了一种潜在社区主题分析模型 LCTA, 该模型基于似然函数生成主题词项与节点的关联图, 并采用期望最大化方法对相关参数进行估计, 在 DBLP 合著关系网络以及 Twitter 社交网络的实

引用格式: 贺超波, 汤庸, 杨阿祧, 等. 基于分布式非负矩阵分解的大规模主题社区挖掘. 中国科学: 信息科学, 2016, 46: 714–728, doi: 10.1360/N112015-00235

© 2016《中国科学》杂志社

验表明 LCTA 可以有效挖掘主题社区; 索勃等 [3] 基于传染病动力学理论, 提出一种动态主题社区发现方法, 该方法通过分析节点间流动的内容信息来发现联系紧密且兴趣相近的主题社区; Xu 等 [4] 针对具有内容属性的复杂网络图, 提出一种基于 Bayes 概率模型的主题社区挖掘方法 BAGC, 可以应用概率推理判断节点的社区归属, 但需要学习多个概率参数; Sachan 等 [5] 结合节点内容以及链接信息对社交网络进行建模的基础上, 运用概率图模型 SNLDA 挖掘主题社区, 该模型不仅可以挖掘节点重叠的社区, 而且社区可以拥有多个主题; 辛宇等 [6] 则针对语义社会网络社区划分的问题提出 SFR 算法, 该算法将语义社会网络的文本内容语义特性和社会关系链接特性相融合, 可有效解决语义社会网络中的主题社区发现问题. 总的来说, 现有主题社区挖掘方法主要关注于如何构建能够有效集成节点链接以及内容信息的挖掘模型, 但忽视了可扩展性方面的优化研究, 如 LCTA、BAGC 以及 SNLDA 等方法都需要划分多个阶段挖掘主题社区, 相关算法可扩展性差, 都不具备大数据量处理能力. 现实世界中的大规模在线社交网络、科技文献合著关系网络等复杂网络不仅节点数量大, 而且节点所关联的内容信息也是海量的, 这对现有主题社区挖掘方法的大数据量处理能力提出了挑战. 针对该问题, 本文提出一种基于分布式非负矩阵分解的主题社区挖掘方法 TCMDNMF (topic community mining based on distributed nonnegative matrix factorization), 该方法的特点在于: (1) 基于联合 NMF 主题社区挖掘模型, 可以统一集成节点链接和内容信息进行社区挖掘, 通过引入 L_1 范数作为稀疏性正则化约束降低了矩阵计算的数据量并提高了模型的计算效率; (2) 采用梯度下降优化方法设计主题社区挖掘模型的迭代求解规则, 并可以基于分布式计算框架 MapReduce 进行实现. 在真实的复杂网络数据集上进行了大量实验, 结果表明本文方法不仅可以保证社区挖掘质量, 并且通过稀疏性正则化约束以及 MapReduce 的优化设计, 提高了计算效率, 具备高度可扩展性, 可以解决现有相关方法无法有效挖掘大规模主题社区的问题.

本文后续内容组织如下: 第 2 节介绍相关工作, 包括 NMF 以及 MapReduce 应用于社区挖掘的情况; 第 3 节给出本文的主题社区挖掘模型、相关求解规则的推导过程和主题社区挖掘算法; 第 4 节详细说明基于 MapReduce 的主题社区挖掘算法的可扩展性设计; 相关实验及结果分析在第 5 节介绍; 第 6 节给出本文工作的总结与展望.

2 相关工作

2.1 NMF 和社区挖掘

由 Lee 和 Seung 提出的 NMF [7] 是一种低秩矩阵近似分解模型, 用于分析非负值矩阵的主要组成部分, 该模型的基本形式为给定一个矩阵 $X \in \mathbb{R}_+^{m\times n}$ 和期望的秩 $k \ll \min(m,n)$, 可以将 X 分解为两个矩阵 $P \in \mathbb{R}_+^{m\times k}$ 和 $Q \in \mathbb{R}_+^{k\times n}$, 且 $X \approx PQ$, 其中 $\mathbb{R}_+$ 为非负值元素集合, P 和 Q 可通过求解函数 $\min J(P,Q)$ 获得, 其中 J 为目标函数, 用于度量 X 和 PQ 之间的误差. NMF 适合对复杂网络数据进行聚类分析以及具有强大的聚类结果可解释能力 [8], 目前已被广泛应用于解决复杂网络中的社区挖掘问题, 这方面具有代表性的工作分别有 Wang 等 [9] 提出一种鲁棒 NMF 方法用于挖掘蛋白质交互网络中的功能模块, 该方法引入 L_1 范数作为正则项降低了噪音数据的影响, 并且采用快速迭代收缩域值策略提高了计算效率; Chan 等 [10] 提出一种基于 NMF 的块模型挖掘社交网络中潜在的社区结构, 可以把社交网络图对应的邻接矩阵近似分解为 3 分解形式: CMC^{T}, 其中 C 表示节点与社区归属关系矩阵, M 表示社区之间的交互程度矩阵, 该模型通过引入约束规则可以有效解决社交网络图过度稀疏以及存在噪音数据的问题, 并可以挖掘重叠社区; Zhang [11] 提出一种基于字典学习的 NMF 社区挖掘方

法, 采用最小二乘误差回归对矩阵分解因子的迭代求解算法进行了改进, 理论分析以及实验结果证明该方法具有线性时间复杂度的优势; 此外, Zhang 等还在文献 [12] 中针对现有基于 NMF 的重叠社区挖掘方法存在节点社区归属关系不确定的问题, 设计了一种基于对称二值 NMF 的重叠社区挖掘模型, 可以更直观地挖掘重叠社区以及发现孤立节点. 总的来说, 现有 NMF-based 社区挖掘方法大都关注基于链接信息的社区挖掘模型以及设计各种矩阵分解优化算法提高单机条件下社区挖掘的效率, 但忽视了节点内容信息的利用以及集群环境下对大规模主题社区的高效挖掘. 本文方法 TCMDNMF 则通过有效融合节点链接和内容信息, 并且运行于分布式集群计算平台 Hadoop, 可以比现有 NMF-based 社区挖掘方法更加有效挖掘大规模主题社区.

2.2 分布式计算框架 MapReduce 和社区挖掘

由 Google 公司提出的分布式计算框架 MapReduce 可以将复杂计算任务划分为多个可并行处理的 Map 和 Reduce 操作, 适合于利用集群计算平台处理大规模数据 [13,14], 目前在大规模复杂网络社区挖掘问题中也已得到广泛关注, 例如, Kang 等 [15] 开发了一种基于 MapReduce 的大图挖掘系统 Pegasus, 该系统包含了用于社区挖掘的谱聚类算法 Heigen, 具有很好的可扩展性; Rytsareva 等 [16] 针对大规模生物图数据的聚类问题设计了一种包含多阶段 MapReduce 处理任务的聚类算法 pClust-mr, 通过并行运行于分布式集群计算平台 Hadoop 显著提高了计算效率; Shi 等 [17] 提出一种基于连接度的预处理和后处理技术 DEPOLD 挖掘大规模社交网络存在的社区结构, 并基于 MapReduce 对核心算法进行了实现, 实验结果证明该技术具有较好的社区挖掘精度以及可扩展性; Moon 等 [18] 针对经典的 GN 社区挖掘方法 [19] 可扩展性差的缺点, 提出了一种改进的最短边介数社区挖掘算法 SPB-MRA, 可以借助 MapReduce 实现所有操作的充分并行化, 从而大大提高了传统 GN 方法的运算性能. Ovelgönne [20] 则实现了一种可并行的基于标签传播的社区挖掘方法, 可以基于 MapReduce 实现并运行于 Hadoop 集群计算平台, 具有高度可扩展性; 此外, Su 等 [21] 提出了一种基于 MapReduce 实现的最大派系枚举方法, 可以用于挖掘社交网络中的重叠社区并显著提升挖掘效率.

对以上相关工作分析可知, NMF 适合用于挖掘复杂网络的社区以及 MapReduce 可以基于分布式集群计算平台解决大规模社区挖掘问题. 本文工作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内容信息的主题社区挖掘模型, 并基于 MapReduce 对挖掘模型的核心算法进行了实现, 有效利用了 NMF 和 MapReduce 的各自优势, 可以高效挖掘大规模复杂网络中的主题社区, 这区别于目前的相关工作.

3 主题社区挖掘模型

复杂网络适合采用图结构表示, 另外节点内容信息可以基于词袋模型进行分析处理, 抽取出每个节点所关联的主题特征词集合, 所以具有节点链接信息和内容信息的复杂网络可以形式化为 $G = (V, E, T)$, 其中 $V = (v_1, v_2, \ldots, v_n)$ 为节点集合, $E = \{e_{ij} | v_i \in V, v_j \in V\}$ 为节点链接边的集合, $T = \{t_1, t_2, t_3, \ldots, t_m\}$ 为主题特征词词典, 且有 $n = |V|$ 以及 $m = |T|$. 不失一般性, G 可以建模为无向无权图, 则其对应的节点链接矩阵可以表示为 $X \in \mathbb{R}_+^{n \times n}$, X 为对称矩阵, 对于任意两个节点 i 和 j, 如果 $e_{ij} \in E$, 那么 X 中对应的元素 $x_{ij} = x_{ji} = 1$, 否则 $x_{ij} = x_{ji} = 0$. 此外, 对于 $\forall v_i \in V$ 都关联一个特征词向量 $[y_{1i}, y_{2i}, \ldots, y_{mi}]$, 其中 y_{ji} 表示节点 v_i 在特征词 t_j 上的取值, 且 $y_{ji} \geqslant 0$, 所有节点的特征词向量可以构成节点特征词关联矩阵 $Y \in \mathbb{R}_+^{m \times n}$. 假设 $C = \{c_1, c_2, \ldots, c_k\}$ 为 G 待挖掘的社区集

合, k 为指定的社区挖掘数, $H \in \mathbb{R}_+^{n\times k}$ 用于表示节点与社区的归属关系矩阵, $S \in \mathbb{R}_+^{k\times k}$ 用于表示社区之间的连接强度矩阵, 根据 NMF 基本模型, X 可以分解为 3 分解形式: $X \approx HSH^{\mathrm{T}}$, 由于 X 为对称矩阵, 那么 S 也为对称矩阵, 令 $H \leftarrow HS^{1/2}$, 则可以进一步简化为 $X \approx HH^{\mathrm{T}}$. 复杂网络的主题社区要求同一社区内的节点链接更加紧密, 内容主题特征更加相同, 可以令基于 X 或 Y 的分解都具有共同分解因子 H. 由于对于 Y 中的任一个列向量可以看作文档主题特征词分布向量, 那么借鉴文献 [22] 基于 NMF 的文档聚类思想, 可得到 $Y \approx WH^{\mathrm{T}}$, 其中 $W \in \mathbb{R}_+^{m\times k}$, 用于表示主题特征词与社区之间的关联强度. 综合以上 X 和 Y 的分解形式, 采用 Frobenius 范数作为误差度量函数, 可以得到基于联合 NMF 的主题社区挖掘模型为

$$\min_{H\geqslant 0, W\geqslant 0}\left\{J(H,W)=\frac{1}{2}(||X-HH^{\mathrm{T}}||_F^2+||Y-WH^{\mathrm{T}}||_F^2+\alpha(||H||_F^2+||W||_F^2))\right\}, \tag{1}$$

其中 $||H||_F^2$ 和 $||W||_F^2$ 为正则项用于提高模型的稳定性, α 为正则化参数, 系数 $1/2$ 用于简化后面的运算. 由于即使主题社区存在重叠的可能性, 某个节点也最多归属少数几个社区, 所以 H 为稀疏矩阵, 同样对于某个主题特征词而言也最多关联少数几个主题社区, 所以 W 也为稀疏矩阵. 在大规模复杂网络中, 节点和主题特征词数量巨大, H 和 W 均为极度稀疏的矩阵, 如果能够控制 H 和 W 的稀疏性, 可以减少数据量, 提高运算效率, 为此引入 L_1 范数对 H 和 W 进行稀疏性约束, 式 (1) 所示的主题社区挖掘模型可以重写为

$$\min_{H\geqslant 0, W\geqslant 0}\left\{J(H,W)=\frac{1}{2}(||X-HH^{\mathrm{T}}||_F^2+||Y-WH^{\mathrm{T}}||_F^2+\alpha(||H||_F^2+||W||_F^2))+\beta(||H||_1+||W||_1)\right\}, \tag{2}$$

其中 β 为稀疏性约束控制参数, 获得目标函数 $J(H,W)$ 的全局最优解是比较困难的, 因为 $J(H,W)$ 对于 H 和 W 是非凸的, 但可以应用数值优化中的梯度下降以及迭代更新规则获得 H 和 W 的局部最优解 [23], 对于 $\forall h_{ik}\in H$ 和 $\forall w_{jk}\in W$, 设计的迭代求解规则分别为 $h_{ik}\leftarrow h_{ik}-\varepsilon_{ik}\nabla_{h_{ik}}J(H,W)$ 和 $w_{jk}\leftarrow w_{jk}-\varepsilon_{jk}\nabla_{w_{jk}}J(H,W)$, 其中 ε_{ik} 和 ε_{jk} 代表梯度下降步长, $\nabla_{h_{ik}}J(H,W)$ 和 $\nabla_{w_{jk}}J(H,W)$ 代表梯度, 可以通过分别计算 $J(H,W)$ 关于 h_{ik} 和 h_{jk} 的偏导数获得, 首先对于 $\nabla_{h_{ik}}J(H,W)$, 有

$$\begin{aligned}
\nabla_{h_{ik}}J(H,W) &= \sum_{a,b}-(x_{ab}-(HH^{\mathrm{T}})_{ab})\frac{\partial(HH^{\mathrm{T}})_{ab}}{\partial h_{ik}}+\sum_{a,b}-(y_{ab}-(WH^{\mathrm{T}})_{ab})\frac{\partial(WH^{\mathrm{T}})_{ab}}{\partial h_{ik}}+\alpha h_{ik}+\beta\\
&= \sum_{a,b}-(x_{ab}-(HH^{\mathrm{T}})_{ab})\frac{\partial h_{ak}h_{bk}}{\partial h_{ik}}+\sum_{a,b}-(y_{ab}-(WH^{\mathrm{T}})_{ab})\frac{\partial w_{ak}h_{bk}}{\partial h_{ik}}+\alpha h_{ik}+\beta\\
&= \sum_{b}((HH^{\mathrm{T}})_{ib}-x_{ib})h_{bk}+\sum_{a}((HH^{\mathrm{T}})_{ai}-x_{ai})h_{ak}\\
&\quad +\sum_{a}((WH^{\mathrm{T}})_{ai}W_{ak}-y_{ai}w_{ak})+\alpha h_{ik}+\beta\\
&= -2(XH)_{ik}+2(HH^{\mathrm{T}}H)_{ik}+(HW^{\mathrm{T}}W)_{ik}-(Y^{\mathrm{T}}W)_{ik}+\alpha h_{ik}+\beta.
\end{aligned} \tag{3}$$

同理可以求得 $\nabla_{w_{jk}}J(H,W)$ 为

$$\nabla_{w_{jk}}J(H,W)=-(YH)_{jk}+(WH^{\mathrm{T}}H)_{jk}+\alpha w_{jk}+\beta. \tag{4}$$

令 $\varepsilon_{ik}=\frac{h_{ik}}{2(HH^{\mathrm{T}}H)_{ik}+(HW^{\mathrm{T}}W)_{ik}+\alpha h_{ik}+\beta}$, $\varepsilon_{jk}=\frac{w_{jk}}{(WH^{\mathrm{T}}H)_{jk}+\alpha w_{jk}+\beta}$, 可以得到 h_{ik} 和 w_{jk} 的迭代更新规则分别为

$$h_{ik}=\frac{(2XH+Y^{\mathrm{T}}W)_{ik}}{(2HH^{\mathrm{T}}H+HW^{\mathrm{T}}W+\alpha H)_{ik}+\beta}h_{ik}, \tag{5}$$

$$w_{jk} = \frac{(YH)_{jk}}{(WH^{\mathrm{T}}H + \alpha W)_{jk} + \beta} w_{jk}. \tag{6}$$

迭代应用规则 5 和 6 可以使目标函数 $J(H,W)$ 收敛并最终可获得 H 和 W 的局部最优解, 由于 $\forall h_{ik} \in H$ 表示节点与社区之间的归属关系强度, 那么通过判断 h_{ik} 的大小可以确定节点 v_i 的归属社区, 具体的基于 NMF 的主题社区挖掘算法描述见算法 1.

算法 1 基于 NMF 的主题社区挖掘算法

Input: 复杂网络节点邻接矩阵 X 和主题特征词关联矩阵 Y, 社区挖掘数 k, 迭代次数 t, 正则化参数 α 和 β, 收敛阈值 φ;

Output: 主题社区挖掘集合 $C = \{c_1, c_2, \ldots, c_k\}$;

1: 随机初始化矩阵 H 和 W;
2: iter $\leftarrow 0$;
3: while (iter $< t$) and $((\|X - HH^{\mathrm{T}}\|_F^2 + \|Y - WH^{\mathrm{T}}\|_F^2) > \varphi)$
4: 分别应用规则 5 和 6 更新 $\forall h_{ik} \in H$ 以及 $\forall w_{jk} \in W$;
5: iter = iter + 1;
6: end while
7: $\forall i = 1, \ldots, k, c_i \leftarrow \emptyset$;
8: for each $v_i \in V$ do
9: $p = \underset{l}{\operatorname{argmax}}\, h_{il}$;
10: $c_p = c_p \cup \{v_i\}$;
11: end for

通过分析以上算法 1 的执行步骤可知, 迭代应用规则 5 和 6 更新 H 和 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. 现实中的复杂网络对应的 X 和 Y 矩阵都是极度稀疏的矩阵, 并且式 (2) 所示的主题社区挖掘模型中加入了对 H 和 W 的稀疏性控制, 因此规则 5 和 6 涉及的矩阵相乘运算都可以采用基于稀疏存储的矩阵乘法进行优化计算. 假设 X 中非 0 元素个数为 μ_X, Y 中非 0 元素个数为 μ_Y, H 和 W 在每一次迭代更新过程中产生的非 0 元素平均个数分别为 μ_H 和 μ_W, 那么可以大致估算 t 次迭代应用规则 5 和 6 的总的时间复杂度为 $O(t(\mu_X + 2\mu_Y + 4\mu_H + 2\mu_W)k)$, 这虽然远小于没有采用稀疏存储优化的运算时间复杂度 $O(t(n^2 + 2mn + 4nk + 2mk)k)$, 但由于大规模复杂网络节点以及主题特征词数量巨大, 并且 H 和 W 在第一次初始化后以及迭代更新求解的开始阶段均为稠密矩阵, 数据量大, 计算复杂度非常高, 依靠单机有限的内存以及 CPU 资源不足于支持算法 1 的稳定以及高效运行, 所以需要考虑提高 H 和 W 迭代更新计算的可扩展性.

4 基于分布式 NMF 的主题社区挖掘算法可扩展性设计

由于规则 5 和 6 主要涉及的是矩阵相乘运算, 矩阵相乘适合采用分布式计算框架 MapReduce 实现, 本文基于 MapReduce 对算法 1 中的迭代更新规则计算进行了重新设计. 考虑到规则 5 和 6 是对称的, 计算方式基本相同, 所以本文只给出规则 5 的 MapReduce 可扩展性实现方法, 具体实现方案如下:

在大规模复杂网络环境中, 规则 5 中的 X 和 Y 均为高维稀疏矩阵, 所有非 0 值元素可以分别采用 $\langle i, j, x_{ij}\rangle$ 和 $\langle i, j, y_{ij}\rangle$ 适合 MapReduce 处理的 key-value 对文件格式进行存储. 此外, 由于 $k \ll \min(m,n)$, 为了在分布式计算环境中实现 H 和 W 参与的矩阵相乘计算充分并行化以及减少数据移动带来的通信开销, 分别把 H 和 W 切割成 n 个和 m 个 k 维行向量组成, 即有 $H = (h_1, h_2, \ldots, h_n)^{\mathrm{T}}$

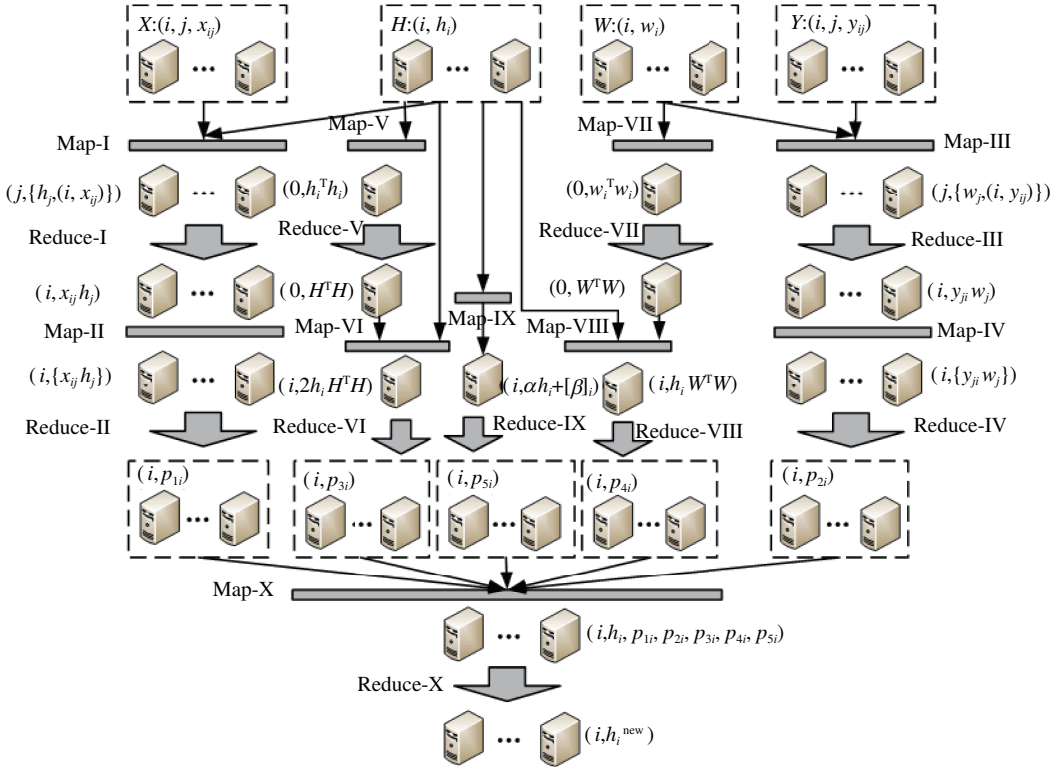


图 1 基于 MapReduce 的 H 矩阵计算流程

Figure 1 Computing matrix H on MapReduce

以及 $W = (w_1, w_2, \ldots, w_m)^{\mathrm{T}}$, 其中 h_i 和 w_j 均为 k 维行向量, 并且在迭代更新过程中均为稀疏向量, 可以只保留非 0 元素, 那么 H 和 W 也可以分别采用 $\langle i, h_i\rangle$ 和 $\langle j, w_j\rangle$ 适合 MapReduce 处理的 key-value 对文件格式进行存储, 从而可以减少数据量. 对规则 5 的计算分解为 6 个部分组成: $P_1 = 2XH$, $P_2 = Y^{\mathrm{T}}W$, $P_3 = 2HH^{\mathrm{T}}H$, $P_4 = HW^{\mathrm{T}}W$, $P_5 = \alpha H + [\beta]$ 以及 $H \leftarrow H \odot (P_1 + P_2) \oslash (P_3 + P_4 + P_5)$, 每一部分的计算可以基于 MapReduce 进行实现, H 矩阵的整体计算流程如图 1 所示, 其中各组成部分的计算过程分别描述如下:

(1) 计算 $P_1 = 2XH$. 令 p_{1i} 为矩阵 P_1 的第 i 个行向量, 那么有 $p_{1i} = \sum_{j=1}^{n} 2x_{ij}h_j$, 则 P_1 的计算可以采用两个 MapReduce 作业完成: Map-I: 读取 X 和 H 的 $\langle j, (i, x_{ij})\rangle$ 和 $\langle j, h_j\rangle$ key-value 对形式的元组, 那么在基于 MapReduce 的分布式集群计算环境中具有相同 key j 的元组将 Shuffle 到相同计算节点, 并生成 $\langle j, \{h_j, (i, x_{ij})\}\rangle$ 形式的元组. Reduce-I: 读取 $\langle j, \{h_j, (i, x_{ij})\}\rangle$ 元组列表, 并输出 $\langle i, x_{ij}h_j\rangle$ 形式的元组. Map-II: 读取 $\langle i, x_{ij}h_j\rangle$, 具有相同 key i 的元组将 Shuffle 到相同计算节点, 并生成 $\langle i, \{x_{ij}h_j\}\rangle$ 形式的元组. Reduce-II: 读取 $\langle i, \{x_{ij}h_j\}\rangle$ 元组列表, 并输出 $\langle i, p_{1i}\rangle$, 其中 $p_{1i} = \sum 2x_{ij}h_j$.

(2) 计算 $P_2 = Y^{\mathrm{T}}W$. 令 p_{2i} 为矩阵 P_2 的第 i 个行向量, 那么有 $p_{2i} = \sum_{j=1}^{m} y_{ji}w_j$, 则 P_2 的计算跟 P_1 一样同样可以采用两个 MapReduce 作业: Map-III, Reduce-III, Map-IV 和 Reduce-IV, 其中各 Map 和 Reduce 操作跟 P_1 的对应操作过程类似.

(3) 计算 $P_3 = 2HH^{\mathrm{T}}H$. 由于 HH^{T} 相乘的结果为 $n \times n$ 的大规模矩阵, 而 $H^{\mathrm{T}}H$ 相乘的结果为 $k \times k$ 的小规模矩阵, 为减少中间计算产生的数据量, P_3 的计算应首先计算 $T' = H^{\mathrm{T}}H$, 然后再计算 $P_3 = 2HT'$. 因为 $T' = H^{\mathrm{T}}H = \sum_{i=1}^{n} h_i^{\mathrm{T}}h_i$, 所以 T' 的计算可以采用 1 个 MapReduce 作业完成;

表 1 数据集特征

Table 1 Dataset characteristics

Dataset	$\|V\|$	$\|E\|$	$\|T\|$	Sparsity(X)	Sparsity(Y)
SCHOLAT	3512	103396	36574	99.2%	99.3%
DBLP10K	10000	55734	100	99.9%	99.9%
CJCA	115634	547263	81235	99.9%	99.9%

Map-V: 读取 $\langle i, h_i\rangle$ 元组并输出 $\langle 0, h_i^{\mathrm{T}} h_i\rangle$, 其中 0 为用于 Shuffle 所有 $h_i^{\mathrm{T}} h_i$ 的 key; Reduce-V: 读入 $\langle 0, h_i^{\mathrm{T}} h_i\rangle$, 计算 $T' = \sum_{i=1}^{n} h_i^{\mathrm{T}} h_i$ 并输出 $\langle i, j, t'_{ij}\rangle$. T' 为小规模矩阵, 则 HT' 的计算可以直接拷贝 T' 到每一个含 h_i 的计算节点并计算 $h_i T'$, 然后汇总输出最终结果, 同样可以采用 1 个 MapReduce 作业完成: Map-VI: 读取 $\langle i, h_i\rangle$, 输出 $\langle i, 2h_i T'\rangle$; Reduce-VI: 读取 $\langle i, 2h_i T'\rangle$, 输出 $\langle i, p_{3i}\rangle$, 其中 $p_{3i} = 2h_i T'$.

(4) 计算 $P_4 = HW^{\mathrm{T}}W$. 同 P_3 的计算一样, 先计算 $T'' = W^{\mathrm{T}}W$, 然后计算 $P_4 = 2HT''$, T'' 和 P_4 的计算都只需要 1 个 MapReduce 作业完成: Map-VII, Reduce-VII, Map-VIII 和 Reduce-VIII, 其中各 Map 和 Reduce 操作跟 P_3 中的对应操作过程类似.

(5) 计算 $P_5 = \alpha H + [\beta]$. P_5 的计算只需要 1 个 MapReduce 作业完成: Map-IX 和 Reduce-IX, 其中 Map-IX 读入 $\langle i, h_i\rangle$, Reduce-IX 输出 $\langle i, \alpha h_i + [\beta]_i\rangle$.

(6) 计算 $H \leftarrow H \odot (P_1 + P_2) \oslash (P_3 + P_4 + P_5)$. $\forall h_i \in H$ 的更新可以采用 1 个 MapReduce 作业进行并行化计算: Map-X 读取 $\langle i, h_i\rangle$, $\langle i, p_{1i}\rangle$, $\langle i, p_{2i}\rangle$, $\langle i, p_{3i}\rangle$, $\langle i, p_{4i}\rangle$ 以及 $\langle i, p_{5i}\rangle$, 具有相同 key i 的元组将 Shuffle 到相同计算节点; Reduce-X 读取 $\langle i, \{h_i, p_{1i}, p_{2i}, p_{3i}, p_{4i}, p_{5i}\}\rangle$, 并输出 $\langle i, h_i^{\mathrm{new}}\rangle$, 其中 $h_i^{\mathrm{new}} = h_i \odot (p_{1i} + p_{2i}) \oslash (p_{3i} + p_{4i} + p_{5i})$.

5 实验分析

5.1 数据集

为验证 TCMDNMF 方法的有效性, 实验采用 3 个典型复杂网络数据集, 第 1 个为社交网络数据集: 学者网 SCHOLAT [1], 第 2 个为进行主题社区挖掘常用的数据集 DBLP10K [4], 第 3 个为科技文献合著关系网络数据集 CJCA. 其中学者网是由作者团队研发的面向科研人员的实名制在线社交网络, 提供学术信息管理、个人主页以及在线社交网络服务, 可以帮助用户建立好友关系网络并分享学术动态和成果, 目前已发展成为国内流行的科研人员协作交流平台. 实验抽取了学者网部分计算机研究领域用户公开的好友关系信息以及学术信息 (包括个人简介、学术论文、科研项目以及知识产权等) 分别作为链接信息和内容信息来源. 科技文献合著关系网络数据集则通过实验开发的 Web 爬虫程序抽取了中国计算机学会会刊, 包括《计算机学报》、《软件学报》以及《计算机研究与发展》等 12 种期刊近十年发表的文章数据, 经过数据预处理后形成. 其中每一个作者代表一个节点, 两个作者存在的合著关系代表一条边, 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主题特征词. 3 个数据集的具体特征信息如表 1 所示, 其中 X 和 Y 的稀疏度计算公式分别为 $\mathrm{Sparsity}(X) = 1 - |E|/(|V| \times |V|)$ 以及 $\mathrm{Sparsity}(Y) = 1 - |\{y_{ij} | y_{ij} \in Y \wedge y_{ij} \neq 0\}|(|V| \times |T|)$. 小规模数据集 SCHOLAT 和 DBLP10K 用于 TCMDNMF 与现有代表性主题社区挖掘方法进行挖掘质量对比分析, 大规模数据集 CJCA 用于测试 TCMDNMF 的可扩展性.

1) http://www.scholat.com.

表 2 学者网社交网络数据集主题社区挖掘质量对比

Table 2 Topic community mining quality comparison on SCHOLAT

Algorithm	k value							
	k=15		k=20		k=25		k=30	
	Density	Entropy	Density	Entropy	Density	Entropy	Density	Entropy
LCTA	0.3947	4.7864	0.3757	5.8463	0.3962	6.5476	0.4038	7.5749
BAGC	0.4015	3.5467	0.3857	3.8546	0.3257	4.5346	0.3619	5.3146
SNLDA	0.3516	5.5476	0.3898	6.7561	0.4086	7.2971	0.3976	8.2846
TCMDNMF	0.4126	2.3745	0.4672	2.7371	0.4263	2.5763	0.4485	2.4718

表 3 DBLP10K 数据集主题社区挖掘质量对比

Table 3 Topic community mining quality comparison on DBLP10K

Algorithm	k value							
	k=50		k=100		k=200		k=300	
	Density	Entropy	Density	Entropy	Density	Entropy	Density	Entropy
LCTA	0.1254	6.3547	0.1154	7.3478	0.1246	8.3547	0.1347	9.2478
BAGC	0.1484	5.5346	0.1673	5.6428	0.1957	5.7987	0.2013	6.2563
SNLDA	0.1348	6.4756	0.1437	6.8794	0.1247	7.6847	0.1304	8.5463
TCMDNMF	0.3268	3.1254	0.3348	3.2345	0.3418	3.3246	0.3312	3.3513

5.2 主题社区挖掘质量评价

对于给定的社区挖掘结果 $C=\{c_1,c_2,\ldots,c_k\}$, 为评价主题社区的挖掘质量, 采用了社区链接密度 Density 和社区主题特征词信息熵 Entropy 作为评价准则, 分别定义如下:

$$\text{Density}(\{c_i\}_{i=1}^k)=\sum_{i=1}^{k}\frac{|\{(v_p,v_q)|v_p,v_q\in c_i,(v_p,v_q)\in E\}|}{|E|},\tag{7}$$

$$\text{Entropy}(\{c_i\}_{i=1}^k)=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{|c_j|}{|V|}\text{entropy}(t_i,c_j),\tag{8}$$

其中 $\text{entropy}(t_i,c_j)=-p_{ij}\log_2 p_{ij}$, p_{ij} 为社区 j 中的节点具有主题特征词 t_i 的比例. 高质量的主题社区挖掘结果不仅社区节点链接紧密, 而且节点主题特征词高度相同, 所以 Density 值越大, 并且 Entropy 值越小, 主题社区挖掘的结果质量就越好. 为验证 TCMDNMF 的社区挖掘质量, 选择 3 种目前具有代表性的主题社区挖掘方法作为基准, 分别为文献 [2] 的 LCTA 方法、文献 [4] 的 BAGC 方法和文献 [5] 的 SNLDA 方法. 实验设置 TCMDNMF 方法的参数 $\alpha=1$, $\beta=0.5$, SCHOLAT 的 $k=15,20,25,30$, t=20, φ=1200, DBLP10K 的 $k=50,100,200,300$, t=30, φ=3000 (经过多次实验测试, 以上两个数据集在以上参数配置下可以获得平滑收敛以及较好的社区挖掘质量). 两个数据集均在设置不同 k 值条件下对比各方法的 Density 和 Entropy 值, 结果如表 2 和 3 所示.

从表 2 和 3 可以看出对于 SCHOLAT 和 DBLP10K 数据集, TCMDNMF 随着 k 值增加, 其 Density 和 Entropy 值不仅优于 LCTA, BAGC 和 SNLDA 方法, 而且具有较好的稳定性, 表明 TCMDNMF 的主题社区挖掘质量可得到较好的保证, 表 4 和 5 则列出了学者网数据集在 $k=30$ 以及 DBLP10K 数

表 4　学者网社交网络典型主题社区

Table 4　Typical topic communities on SCHOLAT

Topic community	Core members	Topic words
Social network	Tang Y (0.6235), Li J G (0.5477), Chen G H (0.4553), Fu C Z (0.4231)	Social network (1.5345), Community mining (0.8547), Friend recommendation (0.6574), Link prediction (0.5328)
Wireless sensor network	Shu L (0.6444), Xiong J B (0.4853), Li R (0.3764), Huang J F (0.3016)	Wireless sensor network (1.3201), Route protocol (0.5632), Data-aware (0.3865)
Data mining	Zhu J (0.5321), Xiao J (0.4021), Lai H J (0.3762), Wang C D (0.3021)	Data mining (1.0212), Cluster (0.6421), Classification (0.3421), Ranking learning (0.2867)
Collaborative computing	Teng S H (0.4983), Liu D N (0.4013), Liang L (0.3014), Zhang W (0.2984)	Collaborative computing (0.8321), Collaborative filtering (0.6321), Collaborative software (0.3211), Collaborative work (0.2876)
Temporal database	Ye X P (0.4132), Liu H (0.3821), Guo H (0.3012), Zuo Y Y (0.2312)	Temporal database (0.9734), Temporal index (0.6312), Temporal XML (0.5321)

表 5　DBLP10K 典型主题社区

Table 5　Typical topic communities on DBLP10K

Topic community	Core members	Topic words
Database	Bancilhon F (0.4521), Biliris A (0.3214), Blakeley J (0.2546), Bruno N (0.2013)	database (1.1124), object (0.8457), sql (0.7458), dbms (0.6364)
Data mining	Agrawal G (0.5564), Bonchi F (0.4785), Giannotti F (0.3647), Han J W (0.3354)	mining (1.3647), patterns (0.9874), frequent (0.7647), sequential (0.6321)
Machine learning	Carin L (0.5843), Dai H H (0.4364), Domingos P (0.3574), Getoor L (0.2357)	models (1.0241), bayesian (0.83247), probabilistic (0.6547), Markov (0.5789)
Information retrieval	Allan J (0.6031), Aslam J (0.4102), Buckley C (0.3654), Clarke C (0.3354)	retrieval (0.9578), information (0.8365), document (0.7027), relevance (0.6874)

据集在 $k = 300$ 时, 分别根据 H 和 W 矩阵获取的典型主题社区及其关联的主题特征词, 括号中的数值代表成员的社区归属强度以及主题特征词的社区关联强度.

5.3　主题社区挖掘可扩展性测试

为验证基于 MapReduce 实现的 TCMDNMF 方法的可扩展性, 采用支持 MapReduce 的开源平台 Hadoop 作为分布式计算环境, 用于测试用的 Hadoop 集群由 10 台具有相同配置的 PC 机组成, 其中一台作为 Master, 其余 9 台作为 Slaver. 每一台机器都具有一个主频为 2.93 GHz 的处理器, 4 GB 内存以及 1 TB 的磁盘存储空间. 由于 TCMDNMF 主要计算消耗都集中在 H 和 W 矩阵的迭代计算, 所以将主要测试分析分布式计算环境中两个矩阵迭代计算方法的可扩展性. 实验采用科技文献合著关系网络数据集 CJCA, 由表 1 可知, 该数据集构建的 X 和 Y 均为极度稀疏矩阵, 通过采用适合 MapReduce 处理的 key-value 对文件格式进行存储, 实际大小分别为 3.7 MB 和 12.6 MB. H 和 W 矩阵在初始化

表 6 不同 k 值下的 H 和 W 矩阵初始化文件数据量

Table 6 H and W matrices initialization file size with different k

Matrix file	k=1000	k=1500	k=2000	k=2500	k=3000
H file (MB)	1072	1608	2141	2677	3213
W file (MB)	752	1128	1504	1879	2255

表 7 不同 k 值下的 H 矩阵第一次迭代更新的计算成本

Table 7 Computation cost of updating H matrix in the first iteration with different k

Component	k=1000		k=1500		k=2000		k=2500		k=3000	
	Shuffle	Time	Shuffle	Time	Shuffle	Time	Shuffle	Time	Shuffle	Time
$2XH$	2642	32.81	4103	59.11	5310	103.71	6624	162.26	7921	225.91
$Y^{\mathrm{T}}W$	2739	44.97	3248	65.34	3785	92.36	4537	126.35	5321	143.62
$2HH^{\mathrm{T}}H$	267	8.76	335	11.38	456	15.21	493	17.28	651	18.76
$HW^{\mathrm{T}}W$	473	12.98	637	16.31	703	21.45	854	22.16	1018	25.82
$\alpha H + [\beta]$	1245	4.68	1713	7.98	2345	12.98	2817	15.12	3487	21.94
H^{new}	875	5.78	1248	6.38	1765	8.35	2058	11.97	2649	15.49

表 8 不同 k 值下的 W 矩阵第一次迭代更新的计算成本

Table 8 Computation cost of updating W matrix in the first iteration with different k

Component	k=1000		k=1500		k=2000		k=2500		k=3000	
	Shuffle	Time	Shuffle	Time	Shuffle	Time	Shuffle	Time	Shuffle	Time
YH	3319	40.75	4583	60.42	6781	142.47	7619	197.55	8978	251.67
$WH^{\mathrm{T}}H$	671	17.18	793	22.31	923	25.13	1247	27.98	1374	31.45
$\alpha W + [\beta]$	1278	6.57	1324	16.12	1698	21.76	2018	24.83	2407	31.47
W^{new}	819	5.67	1345	12.78	1712	15.37	2078	22.37	2349	25.47

后均为稠密矩阵, 表 6 给出了两个矩阵在 k 分别取值为 1000, 1500, 2000, 2500 和 3000 时对应的初始化文件数据量, 可以看出单台 PC 机有限的内存和计算资源难于有效处理 H 和 W 矩阵的大数据量计算, 设置参数 $\alpha = 1$, $\beta = 0.5$, 表 7 和 8 分别给出了在 10 台 PC 组成的 Hadoop 集群计算环境中, 基于 MapReduce 执行 H 和 W 矩阵第一次迭代更新涉及的计算成本, 其中包括各个计算组成部分产生的 Shuffle 数据量 (单位 MB) 和计算任务执行时间 Time (单位 min).

图 2 和 3 分别给出了在不同数量的 PC 组成的 Hadoop 集群计算环境中, 在不同 k 值下基于 MapReduce 执行 H 和 W 矩阵第一次迭代更新需要的处理时间加速比 Speedup 变化曲线, 其中 Speedup $= T_s/T_c$, T_c 是在指定 Hadoop 集群数量为某个值时的运行时间, T_s 为集群数量为 1 时, 即只有一台主机时的运行时间.

从表 7 和 8 可以看出, H 和 W 矩阵一次迭代更新的计算成本集中在 $2XH$、$Y^{\mathrm{T}}W$ 及 YH 等组成部分, 原因在于这些组成部分的计算涉及大量数据分割、数据迁移以及合并等操作, 而 $HH^{\mathrm{T}}H$, $HW^{\mathrm{T}}W$ 以及 $WH^{\mathrm{T}}H$ 等由于优化了计算顺序, 提高了数据处理的本地性以及并发度, 从而降低了计算成本. 另外, 从图 2 和 3 可以看出, H 和 W 矩阵数据量较小时, 加速比稳定, 原因在于数据量较小时, 只充分利用了集群中 2~3 台 PC 机的计算能力, 但随着 k 值的增大, 数据量也不断增多, 不断增加 PC 集群数量加速比可以接近线性增长, 例如当 $k = 2500$ 和 $k = 3000$ 时, 集群数量从 1 台增至 7 台时,

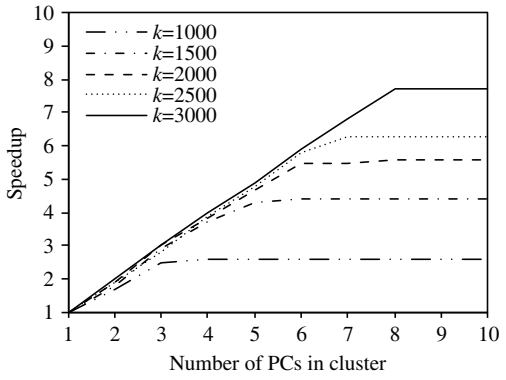


图 2 H 矩阵第一次迭代更新的加速比

Figure 2 Speedup of updating H in the first iteration

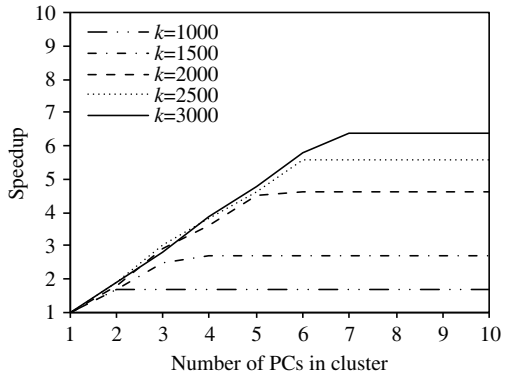


图 3 W 矩阵第一次迭代更新的加速比

Figure 3 Speedup of updating W in the first iteration

表 9 学者网数据集不同 β 值下的稀疏效果对比

Table 9 Sparsity effect comparison with different β on SCHOLAT

β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\triangle$	1642	1531	1432	1356	1204	1298	1359	1643	1784	1901
δ_H	0.23	0.29	0.34	0.37	0.45	0.41	0.38	0.33	0.26	0.25
δ_W	0.19	0.33	0.41	0.47	0.54	0.50	0.46	0.42	0.36	0.32
Time	135	102	86	70	45	56	76	82	109	116

H 和 W 矩阵一次迭代更新的加速比呈近似线性增长. 以上可扩展性测试实验表明基于 MapReduce 实现的 TCMDNMF 方法可充分利用 PC 集群的计算能力提高可扩展性.

5.4 稀疏约束参数 β 的设置分析

式 (2) 所示的主题社区挖掘模型引入了 L_1 范数对 H 和 W 进行稀疏约束, 合理设置 β 值可保证 H 和 W 在每一次迭代更新后将出现更多的 0 元素, 从而 H 和 W 每一次迭代更新的结果文件数据量将不断减少, 每一次迭代计算的耗费时间也将不断减少. 为确保 β 合理取值, 首先在学者网 SCHOLAT 以及 DBLP10K 数据集进行分析实验, 设置 SCHOLAT 参数 $k = 30$, $\alpha = 1$ 以及 $t = 20$, DBLP10K 参数 $k = 300$, $\alpha = 1$ 以及 $t = 30$, 集群数量均为 10, 分别计算 β 取不同值时各数据集迭代结束的误差 $\triangle$、H 的 0 元素个数比例 δ_H、W 的 0 元素个数比例 δ_W 以及总耗费时间 Time (单位 s), 其中 $\triangle = \|X - HH^{\mathrm{T}}\|_F^2 + \|Y - WH^{\mathrm{T}}\|_F^2$, 结果分别如表 9 和 10 所示. 从表 9 和 10 可以看出 $\beta = 0.5$ 时, 误差更小且迭代更新速度更快. 同样, 对于科技文献合著关系网络数据集 CJCA, 设置参数 $k = 3000$, 迭代次数 $t = 50$, β 取值为 0.5, 得到了 H 和 W 每一次迭代更新后的文件数据量 (单位 MB) 以及总的耗费时间 (单位分钟), 结果分别如图 4 和 5 所示.

从图 4 和 5 可以看出, 稀疏性约束参数 β 起到了很好地减少迭代更新数据量以及提高计算效率的作用. 对于本文使用的 SCHOLAT, DBLP10K 和 CJCA 数据集, 经多次测试, 在 $\beta = 0.5$ 时, 可以加快 H 和 W 的稀疏速度, 且迭代收敛效果好于其他取值, 表明针对具体的数据集通过合理地设置 β 值不仅可以保证 TCMDNMF 方法的社区挖掘质量, 而且可以更好地提高该方法的计算效率.

表 10 DBLP10K 数据集不同 β 值下的稀疏效果对比

Table 10 Sparsity effect comparison with different β on DBLP10K

β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\triangle$	3956	3812	3678	3565	3256	3336	3478	3678	3746	3908
δ_H	0.26	0.25	0.32	0.34	0.43	0.42	0.39	0.37	0.27	0.24
δ_W	0.17	0.31	0.43	0.45	0.52	0.49	0.45	0.41	0.35	0.35
Time	202	178	165	149	102	124	151	161	172	186

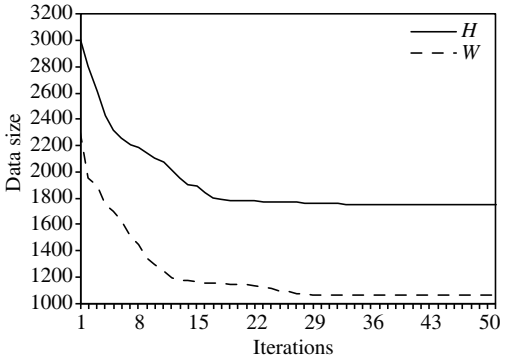


图 4 $\boldsymbol{H}$ 和 $\boldsymbol{W}$ 矩阵迭代更新数据量变化

Figure 4 Data size of H and W after every iteration

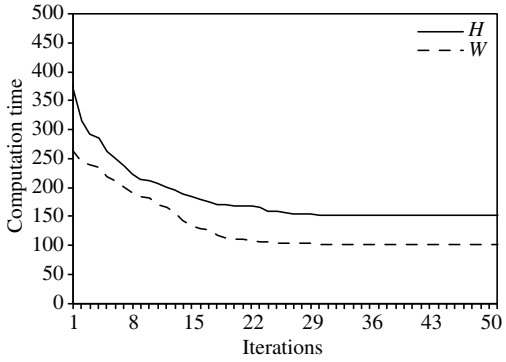


图 5 $\boldsymbol{H}$ 和 $\boldsymbol{W}$ 矩阵迭代更新计算时间变化

Figure 5 Computation time of updating H and W in every iteration

5.5 社区挖掘数 k 的设置分析

在实际实验中, 针对不同规模的复杂网络数据集可以通过不断调整 k 值, 根据 Density 和 Entropy 值确定 k 的最优设置. 对于本文使用的 SCHOLAT, DBLP10K 以及 CJCA 数据集, 首先 α 和 β 均分别设置为 1 和 0.5, 然后 SCHOLAT 的 k 在 [5, 50] 范围内每次递增 5 取值, DBLP10K 在 [50, 500] 范围内每次递增 50 取值, CJCA 则在 [500, 5000] 范围内每次递增 500 取值, 得到的 Density 和 Entropy 值变化趋势如图 6 所示. 通过综合分析 Density 和 Entropy 的变化趋势, 可以看出 SCHOLAT, DBLP10K 以及 CJCA 数据集的 k 值分别在 [15, 30], [50, 300] 和 [1000, 3000] 范围内取值可以获得较好的社区挖掘质量, 这也是以上相关实验设置 k 值的依据.

6 结论

大规模复杂网络主题社区挖掘方法需要具备高度可扩展性, 才能够有效应对大数据量的复杂计算问题. 本文所提出的大规模主题社区挖掘方法 TCMDNMF 基于非负矩阵联合分解模型, 充分利用了非负矩阵分解适合关系网络数据聚类的特点, 有效集成了节点链接和内容信息挖掘主题社区, 此外, 通过引入 L_1 范数进行稀疏性控制降低了矩阵迭代计算过程的数据量以及基于 MapReduce 分布式集群计算框架显著提高了计算效率. 实验证明该方法具有高度可扩展性, 可以高效挖掘大规模主题社区. 由于现实中的大规模复杂网络, 如在线社交网络, 不仅数据量大, 同时也具备动态变化的大数据特点, 在下一步的工作中, 将考虑如何在现有主题社区挖掘模型的基础上研究一种增量式的主题社区挖掘方法,

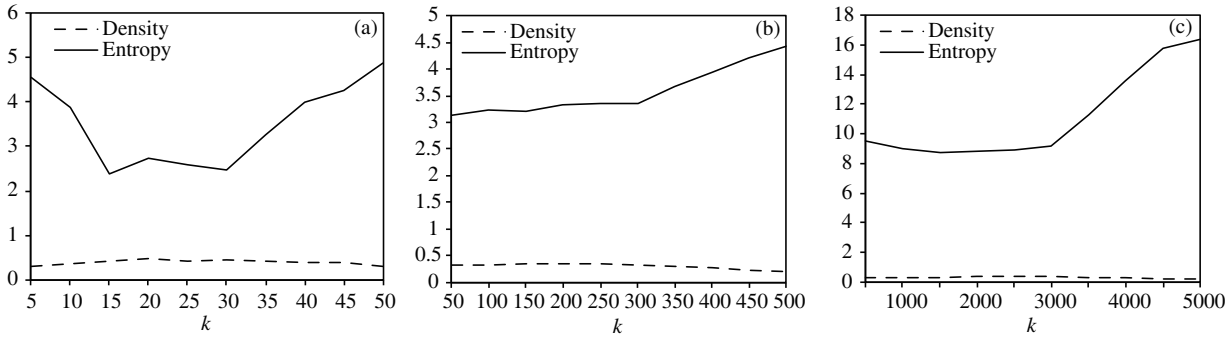


图 6　不同数据集不同 k 值下的 Density 和 Entropy 变化

Figure 6　Density and Entropy with different k on different dataset. (a) SCHOLAT; (b) DBLP10K; (c) CJCA

可以有效应对节点链接以及内容信息动态变化的问题, 方法的实现将更加快速高效地挖掘大规模主题社区. 此外, 由于社区挖掘数 k 的设置更多依靠手工调优的方式获得, 所以研究如何结合具体复杂网络的性质和结构特征自动确定 k 的优化设置也将是下一步的研究方向.

参考文献

1 Zhao Z Y, Feng S Z, Wang Q, et al. Topic oriented community detection through social objects and link analysis in social networks. Knowl-Based Syst, 2012, 26: 164–173

2 Yin Z J, Cao L L, Gu Q Q, et al. Latent community topic analysis: integration of community discovery with topic modeling. ACM TIST, 2012, 3: 63

3 Suo B, Li Z H, Chen Q, et al. Dynamic community detection based on information flow analysis. J Softw, 2014, 25: 547–559 [索勃, 李战怀, 陈群, 等. 基于信息流动分析的动态社区发现方法. 软件学报, 2014, 25: 547–559]

4 Xu Z Q, Ke Y P, Wang Y. A model-based approach to attributed graph clustering. In: Proceedings of the 12th IEEE International Conference on Data Mining, Brussels, 2012. 505–516

5 Sachan M, Dubey A, Srivastava S, et al. Spatial compactness meets topical consistency: jointly modeling links and content for community detection. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, 2014. 503–512

6 Xin Y, Yang J, Xie Z Q. A semantic overlapping community detection algorithm based on semantic data fields. Sci Sin Inform, 2015, 45: 918–933 [辛宇, 杨静, 谢志强. 基于数据场分析的语义重叠社区发现算法. 中国科学: 信息科学, 2015, 45: 918–933]

7 Lee D D, Seung H S. Algorithms for nonnegative matrix factorization. In: Proceedings of the 14th International Conference on Neural Information Processing Systems, Cambridge, 2001. 556–562

8 Ding C, He X F, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of SIAM International Conference on Data Mining, Philadelphia, 2005. 606–610

9 Wang Y Y, Qian X N. Biological network clustering by robust NMF. In: Proceedings of the 5th ACM International Conference on Bioinformatics, Computational Biology and Biomedicine, New York, 2014. 688–689

10 Chan J, Liu W, Kan A, et al. Discovering latent block models in sparse and noisy graphs using non-negative matrix factorization. In: Proceedings of the 22th ACM International Conference on Information and Knowledge Management, New York, 2013. 811–816

11 Zhang Z Y. Community structure detection in social networks based on dictionary learning. Sci China Inf Sci, 2013, 56: 078103

12 Zhang Z Y, Wang Y, Ahn Y Y. Overlapping community detection in complex networks using symmetric binary matrix factorization. Phys Rev E, 2013, 87: 062803

13 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM, 2008, 51: 107–113

14 Dean J, Ghemawat S. MapReduce: a flexible data processing tool. Commun ACM, 2010, 53: 72–77
15 Kang U, Faloutsos C. Big graph mining: algorithms and discoveries. ACM SIGKDD Explor Newsl, 2012, 14: 29–36
16 Rytsareva I, Chapman T, Kalyanaraman A. Parallel algorithms for clustering biological graphs on distributed and shared memory architectures. Int J High Perform Comput Netw, 2014, 7: 241–257
17 Shi J, Xue W, Wang W, et al. Scalable community detection in massive social networks using MapReduce. IBM J Res Dev, 2013, 57: 1–14
18 Moon S, Lee J G, Kang M. Scalable community detection from networks by computing edge betweenness on MapReduce. In: Proceedings of the 1st International Conference on Big Data and Smart Computing, Bangkok, 2014. 145–148
19 Newman M E, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69: 026113
20 Ovelgönne M. Distributed community detection in web-scale networks. In: Proceedings of IEEE/ACM 5th International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, 2013. 66–73
21 Su Y J, Hsu W L, Wun J C. Overlapping community detection with a maximal clique enumeration method in MapReduce. In: Proceeding of the 1st Euro-China Conference on Intelligent Data Analysis and Applications, Shenzhen, 2014. 367–376
22 Xu W, Liu X, Gong Y H. Document clustering based on non-negative matrix factorization. In: Proceeding of the 13th ACM International Conference on Research on Development in Information Retrieval, Toronto, 2003. 267–273
23 Lin C J. Projected gradient methods for nonnegative matrix factorization. J Neural Comput, 2007, 19: 2756–2779

Large-scale topic community mining based on distributed nonnegative matrix factorization

Chaobo HE[1,2], Yong TANG[2*], Atiao YANG[3,2], Gansen ZHAO[2], Hai LIU[2] & Changqin HUANG[2]

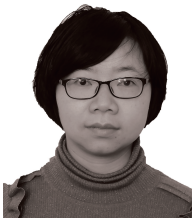
1 *School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou* 510225*, China*;
2 *School of Computer, South China Normal University, Guangzhou* 510631*, China*;
3 *School of Mathematic and Computer Science, Guizhou Normal University, Guiyang* 550001*, China*
*E-mail: ytang4@qq.com

Abstract Mining topic community in complex networks is of great applicable value. However, most of the existing methods cannot effectively mine topic community in large-scale complex networks because of their weak scalabilities. To rectify this problem, we propose a method called TCMDNMF that is based on the joint nonnegative matrix factorization model. The proposed method can effectively integrate node link and content information to mine topic community. We adopt the gradient descent method as the optimized solution to the topic community mining model. Further, to improve the computing efficiency of TCMDNMF, we use L_1 norm as the sparsity regularization term and implement the key algorithms based on the MapReduce distributed computing framework. The results of extensive experiments conducted demonstrate that our method is effective and is highly scalable. Furthermore, it very effectively solves the problem of processing large volumes of data brought by topic community mining in large-scale complex networks.

Keywords complex networks, topic community mining, nonnegative matrix factorization, MapReduce, gradient descent

Chaobo HE was born in 1981. He received a Ph.D. degree in 2014 from South China Normal University, Guangzhou, China. Currently, he is an associate professor at Zhongkai University of Agriculture and Engineering, Guangzhou, China. His research interests are data mining and social computing.

Yong TANG was born in 1964. He received a Ph.D. degree from University of Science and Technology of China. Currently, he is a professor and Ph.D. supervisor at South China Normal University, Guangzhou, China. His research interests include social network service, big data, and temporal databases.



Atiao YANG was born in 1978. She received a Ph.D. degree in 2015 from South China Normal University, Guangzhou, China. Currently, she is an associate professor at Guizhou Normal University, Guiyang, China. Her research interests include social network analysis and temporal knowledge.

Gansen ZHAO was born in 1977. He received a Ph.D. degree from University of Kent, UK in 2007. Currently, he is a professor at South China Normal University, Guangzhou, China. His research interests include trust management and cloud computing.