加权迭代节点匹配算法及其在语言网络中的应用①

张 哲, 宣 琦, 马晓迪, 傅晨波, 俞 立

(浙江工业大学 信息工程学院, 杭州 310023)

摘 要: 复杂网络间节点匹配在很多领域中均具有重要现实意义. 然而, 传统的节点匹配算法通常只利用网络的局部拓扑信息, 在对拥有高对称性的真实网络作用时往往会失效. 为了克服这一缺点, 我们近期利用网络拓扑信息和连边权重信息, 提出了一种新型的同时来计算不同网络间节点相似度的方法, 并在此基础上设计了一种加权迭代节点匹配算法. 将该算法在高度拓扑对称仿真网络对和真实中英文语言网络对上分别进行了测试, 结果表明加权迭代节点匹配算法在此类网络上优于纯拓扑迭代节点匹配算法.

关键词: 复杂网络; 节点匹配; 相似度; 语言网络

Weighted Iterative Node Matching Algorithm and Its Application in Language Networks

ZHANG Zhe, XUAN Qi, MA Xiao-Di, FU Chen-Bo, YU Li

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Node matching between complex networks has practical significance in many areas. However, most of the traditional node matching algorithms which based on the local topological information may lose their efficiencies in many realistic networks, especially in the network with high topological symmetry. In order to overcome this shortage, recently, we proposed a new method to calculate the similarity between two nodes of different networks by utilizing both topological and link-weight information, based on which we designed a weighted iterative node matching algorithm. We test this new algorithm on pairwise artificial networks of high topological symmetry and a pair of real Chinese-English language networks. The results show that the weighted iterative node matching algorithm behaves better than the pure topology based iterative node matching algorithm on this kind of networks.

Key words: complex network; node matching; similarity; language network

随着计算机科学和网络技术的发展, 人们逐渐习惯于用网络来勾画我们所生存的世界, 诸如蛋白质网络[1-3]、语义网络[4-6]、社交网络[7]等. 然而, 由于世间万物的多维度属性, 使得我们可以在不同的表象中观察同一个体. 即相同的个体在不同网络中拥有不同的身份, 而这些不同的网络也由此关联在一起, 形成多层网络[8-10]. 这样的例子包括由同源蛋白质演化成不同生物体内的蛋白质网络之间的关联, 由共同的语义演化出不同语言网络的关联, 以及由共同的人形成不同社交网络之间的关联等. 虽然这些共同的个体在不同的网络中具有不同的身份, 但是它们却具有相同的行为模式. 对这些相同的行为模式进行分析, 可设计网络间节点匹配算法用于发现同源蛋白质, 实现自动语言翻译, 以及社交网络信息过滤等.

目前已有学者利用网络拓扑信息设计了相关的节点匹配算法用于求解该问题[11,12]. 其中纯拓扑迭代节点匹配算法在匹配两个具有较强相关性的无标度网络间对应节点时可获得非常好的效果[13]: 只利用少于2%的已匹配节点对就能正确识别超过 90%的剩余匹配节点. 然而, 这些算法尚存在不足之处, 如该算法在某公司雇员的好友网络和聊天网络这对真实网络中的测试结果并未达到预期目标: 利用超过 10%的已匹配节点对只能正确识别 50%左右的剩余匹配节点对[13]. 会出现这种情况是因为这些传统的算法仅采用网络拓

① 基金项目:国家自然科学基金(61004097,61273232)
收稿时间:2013-12-22;收到修改稿时间:2014-01-23

扑信息, 导致在对很多具有高度拓扑对称性的网络对上进行节点匹配时, 精度不够理想.

事实上, 现实网络通常能提供更多信息, 如连边的权重信息等, 有效利用这些额外的信息将有望进一步提升匹配精度. 我们于近期将网络连边权重信息整合到节点相似度定义中, 并在此基础上设计了加权迭代节点匹配算法[14]. 本文将该算法在高度拓扑对称的仿真网络对和真实中英文语言网络对上进行了测试, 实验结果表明权重迭代节点匹配算法在精度上优于常规迭代节点匹配算法.

文章其余部分安排如下. 第一节提出了仿真模型, 用于创建权重相关网络. 第二节通过整合权重信息扩展了网络间节点相似度定义, 并介绍了加权迭代节点匹配算法. 第三节中用仿真网络和真实网络对算法进行测试, 并对结果进行了相关的分析. 最后一节给出总结和讨论.

1　加权相关网络模型

加权相关网络可分为正相关网络和负相关网络, 其构建主要通过以下三步实现:

1)网络初始化: 根据相同规则生成两个节点数相同的独立网络 $G_1=(V_1,E_1)$ 和 $G_2=(V_2,E_2)$, 其中 $V_i=\{v_1^i,v_2^i,\ldots,v_N^i\}$ 和 E_i 分别表示网络 G_i 的节点集和连边集. 并将两个网络中的所有节点随机一一匹配, 记为 $v_i^1\leftrightarrow v_i^2, i=1,2,\ldots,N$.

2)网络交互: 如果网络 G_1 中某两个节点之间存在连边, 而它们在网络 G_2 中的对应节点之间没有连边, 则以一定的概率 η_1 在 G_2 中对应节点之间增加一条连边, η_1 称为网络 G_1 到网络 G_2 的 “交互度”. 反之, 定义 η_2 为网络 G_2 到网络 G_1 的 “交互度”.

3)权重分配: 分为两种情况. 其一, 如果网络 G_1 (或 G_2)中某两个节点之间存在连边, 且网络 G_2 (或 $\psi(2x,2y)$)中对应节点之间同时也存在连边, 则在 $[1,Q]$ 范围内随机选取一个整数, 记为 $wT+m$, 分配给 G_1 (或 G_2)中的连边, 其中的整数 Q 是所有网络中连边权重的上界. 此时, 如果要构造一对正相关加权网络, 则在区间 $[\pi w,w]$ 内随机选取一个整数, 分配给 G_2 (或 G_1)中对应的连边, 其中 $\pi\in(0,1]$ 为控制关联度的参数; 反之, 如果要构造一对负相关加权网络, 则上述区间改为 $[\pi(Q-w),(Q-w)]$. 其二, 如果网络 G_1 (或 G_2)中某连个节点之间存在连边, 而网络 G_2 (或 G_1)中对应的节点之间不存在连边, 则在 $[1,Q]$ 中随机选取一个整数分配给 G_1 (或 G_2)中的连边, 而保持其在 G_2 (或 G_1)中的对应节点不相连.

为了重现真实网络的高对称性[15], 我们采用 Dorogovtsev, Mendes 和 Samukhin (DMS)所提出的网络模型[16]来生成初始独立网络. 通过该模型生成的网络具有小世界[17], 无标度[18], 高聚类[19,20]等特征. 本文将考虑推广的 DMS 模型: 最开始有 m 个节点彼此相连, 每一步新加入的节点与原有网络中随机选择的一个 m -完全子图相连(如果 m=3 则该完全子图为三角形), 在 T 步以后, 形成一个包含 $T+m$ 个节点以及 C_m^2+mT 条连边的 DMS 网络.

2　迭代节点匹配算法

目前, 我们已经提出了几种基于拓扑结构的节点匹配算法来解决该问题[11-13]. 虽然这些算法在仿真网络上的表现较好, 但在高度拓扑对称性的现实网络中并未达到预期效果. 现实网络大多为加权网络, 充分利用连边权重信息将有望进一步改善现有节点匹配算法的效果. 近期, 我们将连边权重和局部拓扑信息有机结合, 重新设计了不同网络节点之间的相似度公式, 在此基础上提出了一种新颖的加权迭代节点匹配算法[14]. 为完整起见, 我们将用下一小节来简要介绍该相似度计算公式, 并在之后给出具体的算法步骤.

2.1　相似度计算

在基于拓扑信息的节点匹配算法[11,12]中, G_1 中节点 v_i^1 和 G_2 中节点 v_i^2 之间的相似度可以通过下式计算:

$$S(v_i^1,v_j^2)=\frac{n_M(v_i^1,v_j^2)}{n_L(v_i^1)+n_L(v_j^2)-n_M(v_i^1,v_j^2)},\qquad(1)$$

其中, $n_M(v_i^1,v_j^2)$ 表示与节点 v_i^1 和 v_i^2 同时相连的已匹配节点对的数量, 而 $n_L(v_i^1)$ 和 $n_L(v_i^2)$ 则分别表示在网络 G_1 和 G_2 中节点 v_i^1 和 v_i^2 的总邻居数. 公式(1)保证了网络间节点相似度的归一化.

然而, 公式(1)定义的相似度计算公式将无法区分网络中的两个拓扑对称节点, 从而可能会降低整个算法的匹配精度. 为了克服此类不足, 我们在公式(1)中增加连边权重信息, 以此推广相似度的定义. 改进后的节点间权重相似度定义如下:

$$S_w(v_i^1,v_j^2)=\frac{\sum_{l=1}^{k}\left|(w_l^1-w_*^1)(w_l^2-w_*^2)\right|}{\left\|w^1-w_*^1\right\|\cdot\left\|w^2-w_*^2\right\|}S(v_i^1,v_j^2)\,, \quad (2)$$

其中 w_l^1 表示节点 v_i^1 和 v_l^1 之间连边的权重，而 w_l^2 表示节点 v_j^2 和 v_l^2 之间的连边的权重 v_l^1 和 v_l^2 分别是 v_i^1 和 v_j^2 的邻居节点. 这些权重组成向量 $w^1=[w_1^1,w_2^1,\ldots,w_k^1]$ 和 $w^2=[w_1^2,w_2^2,\ldots,w_k^2]$，对应的向量中元素平均值分别记为 w_*^1 和 w_*^2. 值得注意的是，当向量 w^1 或 w^2 中元素均相等时，我们直接定义 $S_w(v_i^1,v_j^2)=S(v_i^1,v_j^2)$. 不难证明公式(2)定义的相似度依旧是归一化的.

2.2 算法过程

同样地，为了文章完整起见，我们在此将简述将权重迭代节点匹配算法[14]的具体步骤. 节点匹配算法是利用若干已匹配节点对提供的信息来匹配网络间的其他节点，假设有 $P_r(P_r<N)$ 对已匹配节点对，则算法的四个步骤描述如下:

1)已匹配节点对选择：仍然采用集中大度值优先策略(centralized large degree priority)(CLDP)来选择已匹配节点对. 即首先在其中一个网络中按以下过程寻找一个节点集 R：①寻找度值最大的节点作为 R 的唯一元素；②寻找与 R 中节点具有最多共同邻居(不包含在 R 中)的节点，将其并入 R 中，直到最终集合 R 包含 P_r 个节点. 然后在另一个网络中找到 P_r 个对应节点，并将这 P_r 对节点作为已匹配节点对. 特别地，我们定义该策略为 CLDP1 或者 CLDP2，如果我们首先从 G_1 或 G_2 中提取节点集 R.

2)相似度计算：网络间节点相似度如公式(2)计算.

3)节点匹配：每次将属于不同网络的一对相似度值最大的未匹配节点对进行匹配，并记为已匹配节点对进行后续运算，即重新转到步骤(2)计算剩余节点之间的相似度.

4)算法终止：当测试集所有节点匹配完，算法终止.

3 算法验证

加权匹配算法的目的是利用已匹配节点对的结构信息和连边权重信息来匹配剩下的节点对，算法的匹配精度定义如下:

$$\varphi=\frac{P_c}{N-P_r}\,, \quad (3)$$

其中 P_r 表示已匹配节点对的数量，N 表示网络中的节点总数，P_c 表示被正确匹配的节点对数.

3.1 仿真实验

首先，利用第一节中引入的加权相关网络模型来生成一对关联网络 G_1 和 G_2. 然后分别用常规迭代节点匹配算法和加权迭代节点匹配算法在 G_1 和 G_2 上进行测试. 此处，两个网络中节点数量设定为相同，即 $N_1=N_2=M=N$. 在每个实验中，共生成 100 组不同的

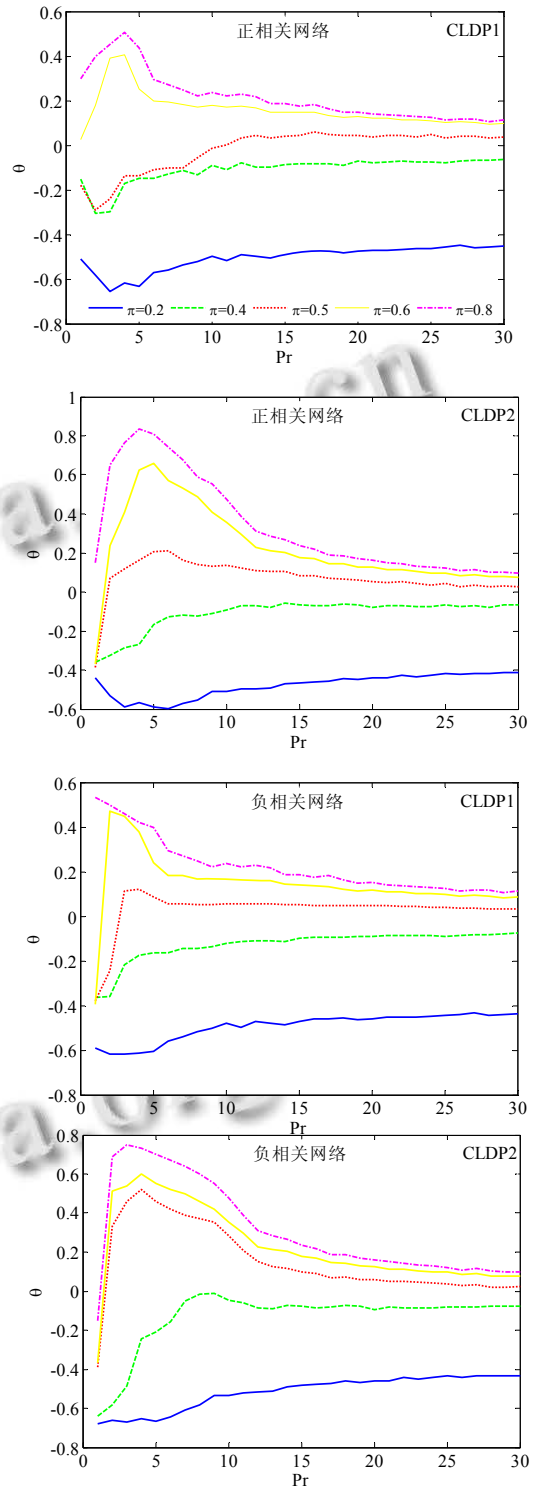

图 3 匹配精度提升图

成对网络. 这些网络对具有共同的参数: 网络的节点数 $N=500$, 交互度 η_1=0.9 、 η_2=0.1, 连边权重的上界 $Q=100$, 网络参数 $m=5$.

为了比较加权迭代节点匹配算法和常规迭代节点匹配算法, 定义如下相对匹配精度提升率:

$$\theta=\frac{\varphi_w-\varphi}{\varphi}, \tag{4}$$

其中 φ_w 和 φ 分别表示加权迭代节点匹配算法和常规迭代节点匹配算法在 100 对网络上的平均匹配精度.

图 3 所示的是权重相关度 π 分别等于 0.2, 0.4, 0.5, 0.6, 0.8 时的匹配精度提升图, 其中前两张图为正相关, 后两张图为负相关. 正如所期望的, 在强关联网络($\pi\geq 0.5$)中, 加权迭代节点匹配算法的表现要好于常规迭代节点匹配算法, 即 $\theta>0$. 而当目标网络 G_1 和 G_2 之间的权重相关性较弱或不存在相关性时, 加权迭代节点匹配算法的表现跟常规迭代节点匹配算法相差不多甚至可能更差. 这可能是由于连边权重信息引入的噪声掩盖了有用的网络拓扑信息, 从而降低了加权迭代节点匹配算法的效果.

3.2 现实网络实验

现实的关联网络对通常比较难以获得, 因为要作测试之用的关联网络对, 不仅要同时获得两个单独的网络, 同时还要得到它们节点之间的匹配关系. 幸运的是, 语言网络的数据相对来说较为容易获得, 一种文字的语段与它的译文可以构成一对相关网络. 我们语言网络来源于 Chiang 等人[21]的 NIST 中文英文翻译工作. 经过预处理, 我们得到待测试的中文网络 G_1 和英文网络 G_2. 此处 G_1 和 G_2 有可能不连通, 通过以下预处理我们获得了两个连通的一一匹配网络对:

1)提取最大团: 分别提取 G_1 和 G_2 的最大团, 记为 $G_1^g=(V_1^g,E_1^g)$ 和 $G_2^g=(V_2^g,E_2^g)$, 其中 V_i^g 和 E_i^g 分别表示 G_i 的节点数和连边数.

2)求节点交集: 选择同时出现在 G_1^g 和 G_2^g 中的节点, 记为 $V_c^g=V_1^g\cap V_2^g$, 分别得到两个子网络 $G_1^c=(V_1^c,E_1^c)$ 和 $G_2^c=(V_2^c,E_2^c)$. 更新 G_1=G_1^c 和 G_2=G_2^c, 如果此时 G_1 和 G_2 均为连通网络, 则预处理过程终止, 否则跳到 1)重复进行.

经过以上处理, 我们最终得到的 G_1 和 G_2 均有 3302 个节点, 并且一一匹配. 此外, 如果两个节点在 G_1 中有连边, 则它们在 G_2 中的匹配节点将有 78.3%的可能性也会有连边, 而从 G_2 到 G_1 的概率则要小得多, 只有 24.8%. 因此此处可以认为 $\eta_1>\eta_2$. 处理后的两个网络的基本统计特征, 包括总节点数 N 、平均度值 $\langle k\rangle$、平均聚类系数 $\langle C\rangle$、平均最短路径 $\langle L\rangle$, 如表 1 所示.

表 1　测试网络的若干统计特征

网络类型	N	$\langle k\rangle$	$\langle C\rangle$	$\langle L\rangle$
中文网络	3302	78.25	0.41	3.55
英文网络	3302	26.98	0.48	2.79

图 4 显示的是 2 个语义网络在 CLDP1 和 CLDP2 策略下的精度提升图. 我们可以看到在两种已匹配节点对选择策略下, 对于不同的已匹配节点对数, 相比常规迭代节点匹配算法而言, 加权迭代节点匹配算法都取得更高的匹配精度, 相对匹配精度提升率均超过了 50%, 提升效果明显.

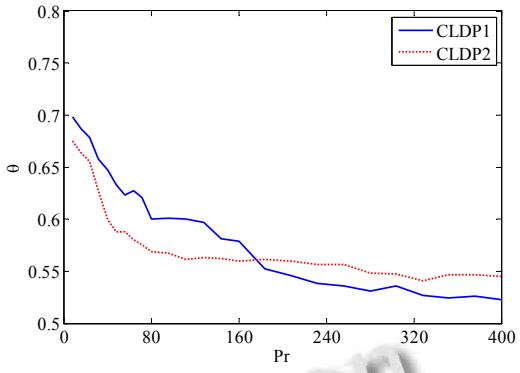

图 4　语言网络匹配精度提升图

4　结论

本文将我们近期提出的加权迭代节点匹配算法在首先在高度拓扑对称仿真网络对上进行了测试, 仿真结果表明在高度拓扑对称性的强关联网络中加权迭代节点匹配算法的表现优于常规迭代节点匹配算法, 更具实用性. 同时, 我们将该算法用于匹配现实的中英文语言网络, 实现机器自动匹配中英文词汇, 结果表明: 相比传统迭代节点匹配算法, 新的加权迭代节点匹配算法取得了更好的精度, 在此例子上相对匹配精度提升率均超过 50%, 提升效果明显.

参考文献

1 Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science, 2002, 297(5586): 1551−1555.

2 Barabási AL, Oltvai ZN. Network biology: understanding the

cell's functional organization. Nature Reviews Genetics, 2004, 5(2): 101−113.

3 殷志祥.蛋白质结构预测方法的研究进展.计算机工程与应用,2004,20:54−57.

4 李蕾,王楠,钟义信,郭祥昊,韩鹏,贾自燕,高清霞.基于语义网络的概念检索研究与实现.情报学报,2000,5:525-531.

5 Cancho RF, Solé RV, Khler R. Patterns in syntactic dependency networks. Physical Review E, 2004, 69(5): 051915.

6 Amancio DR, Antiqueira L, Pardo TA, da F. Costa LC, Oliveira Jr ON, Nunes MG. Complex networks analysis of manual and machine translations. International Journal of Modern Physics C, 2008, 19(04): 583−598.

7 吴增海.社交网络模型的研究.合肥:中国科学技术大学, 2012.

8 Kurant M, Thiran P. Layered complex networks. Physical review letters, 2006, 96(13): 138701

9 Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S. Catastrophic cascade of failures in interdependent networks. Nature, 2010, 464(7291): 1025−1028.

10 Xuan Q, Du F, Wu TJ. Empirical analysis of Internet telephone network: From user ID to phone. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2009, 19(2): 023101-023101-10.

11 Xuan Q, Wu TJ. Node matching between complex networks. Physical Review E, 2009, 80(2): 026103.

12 Xuan Q, Du F, Wu TJ. Iterative node matching between complex networks. Journal of Physics A: Mathematical and Theoretical, 2010, 43(39): 395002.

13 Du F, Xuan Q, Wu TJ. One-to-many node matching between complex networks. Advances in Complex Systems, 2010, 13(6): 725−739.

14 Xuan Q, Zhang Z, Fu C, Gharehyazie M, Yu L. Node matching between weighted networks. Advances in Complex Systems, preprint.

15 Xiao Y, Xiong M, Wang W. Emergence of symmetry in complex networks. Physical Review E, 2008, 77(6): 066108.

16 Dorogovtsev SN, Mendes JFF, Samukhin AN. Size-dependent degree distribution of a scale-free growing network. Physical Review E, 2001, 63(6): 062101.

17 Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’networks. nature, 1998, 393(6684): 440−442.

18 Barabási AL, Albert R. Emergence of scaling in random networks. science, 1999, 286(5439): 509−512.

19 Ravasz E, Barabási AL. Hierarchical organization in complex networks. Physical Review E,2003,67(2): 026112.

20 杨博,刘大有,金弟,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66.

21 Chiang D. A hierarchical phrase-based model for statistical machine translation. Association for Computational Linguistics, 2005: 263−270.