

差异化社会网络节点角色的链路预测模型

伍杰华^{1,2}

(1 广东工贸职业技术学院 计算机工程系, 广东 广州 510510;

2 华南理工大学 信息科学与技术学院, 广东 广州 510641)

摘 要: 以经典的 Common Neighbor 算法为例, 提出了一种基于社区划分的差异化节点角色的链路预测模型, 该模型首先采用 Clauset-Newman-Moore 算法挖掘社会网络结构属性, 同时引入节点连接度和社区整体参与度的定义, 差分处理社区内外邻接节点和不同社区的贡献, 采用有监督的学习训练方法分别对社区内节点对和社区间节点对进行链路预测. 人工网络和真实网络中的实验证明, 该模型能够提高基于相似度算法对节点对链路预测的准确率, 并为该类模型的研究提供一种新的方案.

关键词: 链路预测; 社会网络; 节点角色; 社区划分; 共同邻接节点

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)10-0050-04

Link Prediction in Social Network Via Differentiating the Role of Vertexes

WU Jie-hua^{1,2}

(1 Department of Computer Science and Engineering, Guangdong College of Industry and Commerce,
Guangzhou 510510, China; 2 College of Information Science and Technology,
South China University of Technology University, Guangzhou 510641, China)

Abstract: This paper proposes a new link prediction model basing on community partition to differentiate diverse contribution of different vertexes which taking Common Neighbor algorithms as examples. Firstly, the model uses Clauset-Newman-Moore algorithm to partition community and generates pairs via a different aspect, then applies different weight to inner-community and outer-community neighbor vertexes, finally provides a supervised training method to infer the missing links. The experimental results demonstrate that the algorithm can not only raises prediction precision but also provides a kind of effect method for research of similarity link predication.

Key words: link prediction; social network; node role; community partition; common neighbor

1 引言

社会网络的不断涌现是近十年来互联网领域最令人瞩目的事件, 诸如 Twitter, LinkedIn 和 Facebook 等公司的不断发展使人们对深入挖掘社会网络中的数据和模式产生浓厚的兴趣, 链路预测(Link Prediction)便是其中的重点研究领域.

网络用链路将各个孤立的节点相连在一起, 组成一种网状的数据结构. 链路预测是利用基于网络

历史结构信息预测未来节点之间产生的链路的可能性. 该类模型在众多领域有着广泛的应用, 例如预测社交网络中的朋友关系, 预测交易网络中哪些人将会产生交易, 检测复杂生物信息如蛋白质交互信息, 发现恐怖分子网络中的恐怖信息等.

本文提出了一种基于社区划分^[1]的差异化节点角色的链路预测模型. 首先介绍链路预测模型的框架, 提出了一个基于不同社区尺度的监督学习预测框架. 在框架中, 把链接节点对分成两类: 社区内节

收稿日期: 2012-12-20; 修回日期: 2013-02-07

基金项目: 国家自然科学基金项目(61003045)

点对和社区间节点对, 针对不同类别采取不同的算法, 同时把该框架结合经典的 CM 和 Jaccard 公式进行实验, 从而把该方法推广到所有基于相似度的链路预测模型中。

$$\text{CM: } \Gamma(v_i) \cap \Gamma(v_j) \quad (1)$$

$$\text{Jaccard: } \frac{\Gamma(v_i) \cap \Gamma(v_j)}{\Gamma(v_i) \cup \Gamma(v_j)} \quad (2)$$

贡献主要有以下三点:

(1) 创造性提出了采用社区划分挖掘社区内部结构属性并应用与链接预测模型中。

(2) 在社区划分基础上, 提出基于不同节点贡献的链路预测模型, 模型中从属不同社区的节点对链路预测有不同的贡献。

(3) 提出了基于局部和全局的链路预测概念, 同时预测模型能够应用到所有基于邻接节点计算相似度的链接预测算法。

2 链路预测模型

2.1 社区发现和模型定义

定义 1 $G = (V, E)$ 定义网络结构, 其中 $N = \{n_1, n_2, n_3, \dots, n_m\}$ 和 $E = \{(v_i, v_j) \mid v_i, v_j_connect\}$ 分别定义网络中节点和链接的集合, 其中 (v_i, v_j) 定义为节点对, $\Gamma(v_i, v_j)$ 为节点对的邻接节点集合。

定义 2 划分社区结果为: $C = \{C_1, C_2, C_3, \dots, C_n\}$, 每一个 C_i 都有其包含的节点和链接。

定义 3 $S(a, b)$ 定义 a, b 两个节点之间的相似度, 其中 $S_{\text{local}}(a, b)$ 和 $S_{\text{between}}(a, b)$ 分别对应社区内和社区间的节点对。

由社区定义可知: 社区内部链接集中而社区之间链接相对稀疏, 同时不同社区的规模也不一致。针对这一点, 本文提出了一种新颖的采用社区划分差异化节点角色算法处理社区内外部邻接节点, 并采用由 Clauset-Newman-Moore^[2] 提出来的社区划分算法(CNM 算法), 该算法通过移除网络中的边来将整个网络划分成为合适的社团结构, 划分结果中不存在社区重叠的节点, 同时任意的节点只属于一个社区, 链接则可以连接同一个社区内的节点和不同社区之间的节点。采用社区划分的好处在于深入挖掘网络内部结构属性, 并作为后续差异化邻接节点角色的基础, 最后采用 CM 和 Jaccard 两个基于邻接节点的模型作为验证手段, 为了简化计算, 下文采用 CM 算法为例进行建模。

2.2 社区内部链路预测

社区内部节点对链路预测的问题描述为: 给定

节点 a 属于社区 C_i , 节点 b 属于社区 C_i , 计算 $S(a, b)$ 。由于社区具备相同的性质和兴趣, 所以社区内部节点之间的属性也较近似, 社区内部链路预测算法基于同一个社区之间两个节点的链接可能受社区内部环境影响比社区外部环境影响大, 以及在社区内部连接度高和连接度低的邻接节点对预测的贡献不一致, 其主要思想是提高社区内部邻接节点的贡献权重, 按照邻接节点的连接度区分其贡献, 同时降低社区外部邻接节点的贡献权重, 用 $S_{\text{in}}(a, b)$ 和 $S_{\text{out}}(a, b)$ 分别代表社区内外部环境对该节点对的影响, 那么:

$$S_{\text{local}}(a, b) = S_{\text{in}}(a, b) + S_{\text{out}}(a, b) \quad (3)$$

为了实现这一点, 这里需要解决两个问题: 如何定义邻接节点角色并区分和其所处同一社区内部邻接节点的贡献而提高权重; 如何定义获取外部社区邻接节点的贡献并降低权重。由于预测节点对的邻接节点之间的连接方式多样化, 可能在同一社区内或者在外部社区, 直接用出入度表示其贡献权重不太适合, 本文通过定义社区内部节点连接度和社区参与度来差异化邻接节点角色。

首先用 D_i 表示节点 i 与其所在社区 C_i 其他邻接节点之间出入度, $\bar{D}_i$ 表示社区 C_i 所有节点出入度的均值, σ_i 表示标准差, 则邻接节点 i 在社区 C_i 内部的连接度为

$$k_i = (D_i - \bar{D}_i) / \sigma_i \quad (4)$$

式中, k_i 代表某个节点和其社区内部节点的连接度, 连接度越大的节点, 其对 $S(a, b)$ 的贡献度越大。

由于邻接节点社区外部的节点进行连接, 我们还需要定义参与系数来刻画该节点与其他社区邻接节点的链接情况并差异化处理不同社区对不同节点的贡献, 所以节点 i 的参与系数定义 p_i 为

$$p_i = 1 - \sum_{k=1}^{N_c} \left(\frac{D_k}{D_i} \right)^2 \quad (5)$$

式中, D_k 描述节点 i 与社区 C_i 中节点的连接数, D_i 描述节点 i 的总连接数, N_c 是社区的数目, 从上式还可以得出:

$$p_k = 1 - \left(\frac{D_k}{D_i} \right)^2 \quad (6)$$

式中, p_k 指节点 i 与某一特定社区 C 的参与系数, 如果是 0 的则代表其所有的邻接节点均在其所处社区的内部。

本文根据 k_i 和 p_i 对 C_i 内部节点的角色进行划分, 如果 k_i 和 p_i 值较大或者 k_i 较大而 p_i 趋向于 0, 那么归结为 hub 节点^[3], 其余都归结为普通节点, 在

社会网络中, hub 节点是网络中不同社区的核心枢纽节点, 如果两个节点的邻接节点都有 hub 节点, 那么其链接的可能性相对会更高. 所以, 用 $S_{in}(a, b)$ 是社区内部 hub 节点的贡献权重和社区内部普通节点的贡献权重相加, 其公式表示为

$$S_{in}(a, b) = S_{in-hub}(a, b) + S_{in-ordinary}(a, b) \quad (7)$$

式中, hub 节点的贡献权重 $S_{in-hub}(a, b)$ 公式为

$$S_{in-hub}(a, b) = \sum_{i \in \Gamma(a, b) \in \text{hub}}^{N_{\text{hub}}} \delta k_i \quad (8)$$

式中, N_{hub} 指所有 $x \in \Gamma(a, b)$ 且为 hub 节点个数, k_i 是第 i 个 hub 节点的连接度, δ 是参数.

同时为了区别各普通节点的影响并简化计算复杂度, 我们把节点按照连接度分成 r 个等级, 计算第 i 个等级的连接度均值 lk_i , 并用其表示该等级所有节点的连接度. 普通节点的贡献 $S_{in-ordinary}(a, b)$ 公式为

$$S_{in-ordinary}(a, b) = \sum_{i=1}^r (1 - \delta) lk_i \times N_{lk_i} \quad (9)$$

式中, N_{lk_i} 指所有 $x \in \Gamma(a, b)$ 且属于第 i 个等级节点的数目.

相反, 社区外部的节点对 $S(a, b)$ 的影响相对较弱, 所以计算社区外部节点对 a, b 所在社区的参与度:

$$S_{out}(a, b) = \sum_{x \in \Gamma(a, b) \cup x \in C_i}^{N_c} \sum_{j=1}^{\text{count}(x \in C_i)} k_j (1 - p_{k_i}) \quad (10)$$

2.3 社区之间链路预测

社区之间的连接描述为: 给定 a 属于社区 C_A , b 属于社区 C_B , 计算 $S_{\text{between}}(a, b)$, 其中:

$$S_{\text{between}}(a, b) = S_{b_in}(a, b) + S_{b_out}(a, b) \quad (11)$$

该算法基于思想: 在社会网络中不同社区的权重是不一致的并且如果 C_A 与 C_B 之间联系越紧密, 那么 a, b 之间连接程度的可能越高. 由此, 我们需要量化区分社区的属性及定义其贡献度, 用 $E_{A, B}$ 表示 C_A 和 C_B 之间存在的边连接数, 用 $\bar{E}_{A, B}$ 表示 C_A 和 C_B 之间存在可能的边连接数, $R(C_A, C_B)$ 表示 C_A 和 C_B 的相关度, 公式如下所示:

$$R(C_A, C_B) = E_{A, B} / \bar{E}_{A, B} \quad (12)$$

并通过式(7)获得两社区内部节点的链接度贡献, 可得以下公式:

$$S_{b_in}(a, b) = (1 + R(C_A, C_B)) \times S_{in}(a, b) \quad (13)$$

其中为所有 $x \in \Gamma(a, b)$ 且 $x \in C_A \wedge x \in C_B$ 的集合. 对于在社区之外的节点, 需要考虑该节点所在社区的贡献, 本文提出用社区间连接度标示社区的

贡献, 用 D_{C_i} 表示社区 C_i 与其他社区的出入度, $\bar{D}_{C_i}$ 表示网络中所有社区出入度的均值, σ_G 表示标准差, 社区 C_i 与其他社区的连接度:

$$k_{C_i} = (D_{C_i} - \bar{D}_{C_i}) / \sigma_G \quad (14)$$

式中, N_{C_i} 为所有 $x \in \Gamma(a, b)$ 且 $x \in C_i$ 的集合, 并通过式(10)获得其余社区节点的连接度和贡献度, 可得公式为

$$S_{b_out}(a, b) = \sum_{i=1}^{N_c} (R(C_A, C_i) + R(C_B, C_i)) \times S_{out}(a, b) \times k_{C_i} \quad (15)$$

2.4 算法框架

Algorithm Process:

Given: $G = (V, E), C = \{C_1, C_2, C_3, \dots, C_n\}$

Result: $S(v_i, v_j)$ and $CN(v_i, v_j)$ ($v_i, v_j \notin E$)

Initialize $k_i = D_i - \bar{D}_i / \sigma_i$ // 获取各节点的社区内连接度

$$p_i = 1 - \sum_{i=1}^{N_c} \left(\frac{D_i}{\bar{D}_i} \right)^2 \quad p_{i^*} = 1 - \left(\frac{D_{i^*}}{\bar{D}_{i^*}} \right)^2 // \text{分别获取社区内}$$

外部的贡献度

$R(C_A, C_B) = E_{A, B} / \bar{E}_{A, B} \quad k_{C_i} = D_{C_i} - \bar{D}_{C_i} / \sigma_G$ // 计算社区间相关度和连接度

Begin:

for each pair(v_i, v_j) $\notin E$

if $v_i, v_j \in C_k$

$$S_{in}(v_i, v_j) = \sum_{a \in \Gamma(v_i, v_j) \in \text{hub}}^{N_{\text{hub}}} \delta k_a + \sum_{m=1}^r \beta \times lk_m \times N_{lk_m}$$

$$S_{out}(v_i, v_j) = \sum_{x \in \Gamma(v_i, v_j), C_i}^{N_c} \sum_{j=1}^{\text{count}(x \in C_i)} k_j (1 - p_{k_i})$$

$$S_{\text{local}}(a, b) = S_{in}(a, b) + S_{out}(a, b)$$

else if $v_i, v_j \notin C_k$

$$S_{b_in}(v_i, v_j) = (1 + R(C_A, C_B)) \times S_{in}(v_i, v_j)$$

$$S_{b_out}(v_i, v_j) = \sum_{i=1}^{N_c} (R(C_A, C_i) + R(C_B, C_i)) \times S_{out}(v_i, v_j) \times k_{C_i}$$

$$S_{\text{between}}(v_i, v_j) = S_{b_in}(v_i, v_j) + S_{b_out}(v_i, v_j)$$

endif

endfor

3 实验分析

3.1 实验方法和评价

本文分别对人工生成网络和现实网络开展实验, 给定一个 $G = (V, E)$, 随机移除 E 中 10% 的边并记为预测子集 E_r , 其他 90% 的边作为训练数据集 $E = E_r \cup E_t$, 则 $\emptyset = E_r \cap E_t$, 节点总数保持不变, 进行原始 CM、Jaccard 和改进后的算法(简称 CM_E 和 Jaccard_E)比较, 取 δ 值为 0.6, 把社区内部、社区之间和全局预测节点作为实验对象, 采用 10—

folder 交叉验证方法,交叉重复验证 10 次,每次选择一个子集作为测试集,并将 10 次的平均交叉验证识别正确率(precision)作为评价指标:

precision=
$$\frac{\text{count}(S(v_i, v_j)) + \text{count}(\text{CN}(v_i, v_j))}{\text{count}(E_r)} \tag{16}$$

3.2 实验数据集

为了评价算法差分化节点角色的能力,分别使用了人工网络和真实网络开展实验. 人工网络采用 Babarasi^[4] 提出的随机网络生成算法,生成具备不同属性的 5 个人工网络,真实网络采用 email^[5] 和 port^[6] 数据集进行实验. 具体结构属性如表 1、表 2 所示.

表 1 真实网络结构属性

Network	V	E	C_Num	Modularity
email	1 116	4 906	19	0. 507
port	327	1 913	8	0. 324

表 2 人工网络结构属性

BA	V	E	C	AD	C_Num
BA_1	30	210	0. 573	14	4
BA_2	80	1 083	0. 576	27. 07	5
BA_3	100	1 212	0. 432	24. 24	4
BA_4	200	4 334	0. 476	43. 34	3
BA_5	300	7 220	0. 392	48. 13	4

其中 V,E,C,AD 分别表示网络的节点,边、聚类系数和平均节点度,C_num 和 Modularity 分别表示网络划分社区的数目和其模块度.

3.3 链路预测评价

从表 3、图 1 和 2 中可以看出,差分化处理节点贡献后的模型对三类实验对象的预测精确度都有所提高,这是由于对社区内部的邻接节点和社区外部的节点赋予了不同的贡献权重并引入了社区贡献度进行差分化处理邻接节点,但是对于 BA_3,其识别效果相对不太明显,这是由于该网络结构的聚类系数较低,社区连接紧密度稀疏^[7]所致.

表 3 人工网络全局预测结果

算法	BA_1	BA_2	BA_3	BA_4	BA_5
CM	0. 324	0. 413	0. 331	0. 683	0. 866
CM_E	0. 342	0. 417	0. 332	0. 689	0. 874
Jaccard	0. 307	0. 414	0. 323	0. 684	0. 857
Jaccard_E	0. 322	0. 422	0. 330	0. 688	0. 865

3.4 真实网络链路预测评价

真实网络的实验结果也证明本文框架的有效

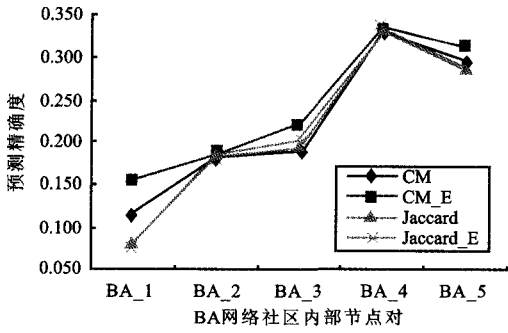


图 1 人工网络社区内部预测

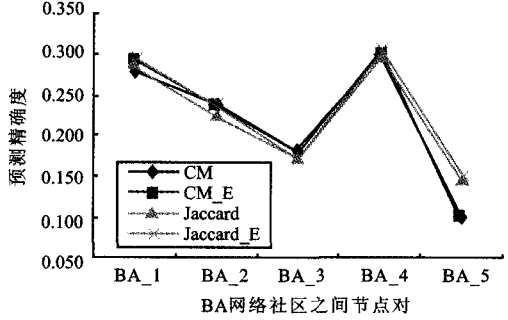


图 2 人工网络社区之间预测

性. 尤其对于 email 网络,社区内部比社区之间节点对的预测差异化更加明显,这是由于真实网络划分的社区数目较多,社区间节点数目差异大,部分社区只有几个节点,所以导致节点数目较少的社区贡献较难量化处理,这也是后面需要进一步提高的工作.

4 结束语

采用基于社区划分并差分化社会网络中节点角色的链接预测模型可以提高社会网络中链路预测的精确度,有效解决量化邻接节点贡献权重的问题. 实验证明,其相比一般的基于相似度的算法更加优越,也能够推广到其他类似的模型当中.

但是,如果社区划分后不同社区结构差异化较大的时候,得到的预测结果就不太理想,所以,采用更优的社区划分算法和选择更有效的社区属性进行链路预测是下一步的研究工作.

参考文献:

[1] Juszczyszyn K, Musial K, Budka M. Link prediction based on subgraph evolution in dynamic social networks[C]// 3rd IEEE International Conference on Social Computing. Boston, MA, USA, 2011.

[2] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical review E, 2004, 70(6): 066111.

法,本文算法耗时最短. 由于没有进行 8 种等距变换,运算速度有很大改善,运算时间大约为基本算法的 1/60,文献[8]算法的 1/18.

压缩比如表 3 所示. 文献[8]算法最高,本文算法次之,基本算法最低.

表 3 压缩比较		kB			
		基本 算法	文献[8] 算法	JPEG 压缩	本文 算法
Lena	原始图像	64	64	64	64
	压缩后图像	12.5	2.94	10.5	7.0
	CR	5.12	21.78	6.09	9.11
Cameraman	原始图像	64	64	64	64
	压缩后图像	12.5	2.94	10.4	5.8
	CR	5.12	21.78	6.15	10.99
Couple	原始图像	64	64	64	64
	压缩后图像	12.5	2.94	11.5	7.0
	CR	5.12	21.78	5.57	9.14

6 结束语

将小波变换和分形编码相结合对图像进行压缩编码,小波分解形成三个方向的子带,自然的对分形编码中的定义域块和值域块进行了分类,从而减少了编码搜索的时间. 同时,小波分解产生的各级子带间具有较强的相似性,分形的优势正好在于自相似性强的图像,如何利用分形开采这种相似性,提高图

像解码质量,还可以进行深入的研究.

参考文献:

[1] 夏良正,李久贤. 数字图像处理[M]. 2 版. 南京:东南大学出版社,2006.

[2] Barnsley M. Fractal everywhere[M]. San Diego, CA: Academic Press, 1988.

[3] Jacquin A. A fractal theory of iterated markov operators with application to digital Image coding[D]. USA, Atlanta: Georgia Institute of Technology, 1989.

[4] Roberto Rinaldo, Giancarlo Calvagno. Image coding by block prediction of multiresolution subimages[J]. IEEE Trans. on Image Processing, 1995, 4(7): 909-920.

[5] Geoffrey M Davis. A wavelet-based analysis of fractal image compression[J]. IEEE Trans. on Image Processing, 1998,7(2):141-154.

[6] 谢永华,傅德胜,任卫华. 一种基于小波变换的图象分形编码压缩算法的研究[J]. 中国图象图形学报, 2003, 8(7):839-842.

[7] 张梁斌,奚李峰. 一种基于小波域的分形编码预测方法[J]. 计算机应用与软件, 2007, 24(6):68-70.

[8] 娄莉. 一种基于小波域的分形图像编码改进算法[J]. 计算机技术与应用,2010(7):134-136.

[9] 张宗念,马义德,余英林. 基于方向性零树小波的分形图像编码[J]. 电子科学学刊,2000,22(5):780-784.

作者简介:

牟宇飞 男,(1980-),硕士研究生. 研究方向为图像压缩、加密.

(上接第 53 页)

[3] Lü L,Zhou T. Link prediction in weighted networks: the role of weak ties[J]. EPL(Europhysics Letters), 2010,89(1):18001.

[4] Barabási, Albert-László,Réka Albert. Emergence of scaling in random networks[J]. Science, 1999 (286): 509-512.

[5] Arenas A, Cabrales A, Díaz-Guilera, et al. Search and congestion in complex networks[C]// Statistical mechanics of complex networks. Berlin, Heidelberg: Springer, 2003:175-194.

[6] Kunegis J. KONECT: the koblenz network collection [C]// Proceedings of the 22nd international conference

on World Wide Web companion. Brazil: International World Wide Web Conferences Steering Committee, 2013:1343-1350.

作者简介:



伍杰华 男,(1982-),博士,讲师,高级工程师. 研究方向为数据挖掘、机器学习、社会网络分析.