

基于能量函数和模块最优化的不确定图聚类*

丁悦¹, 张阳^{1,2}, 王勇³, 李伟卫¹

(1. 西北农林科技大学 信息工程学院, 陕西 杨凌 712100; 2. 南京大学 计算机软件新技术国家重点实验室, 南京 210093; 3. 西北工业大学 计算机学院, 西安 710072)

摘要: 为证明不确定性的存在对聚类结果不可忽略的影响, 改进了基于能量模型布局 and 模块化聚类的算法 LinLogLayout, 使之可以处理不确定图数据。提出了不确定图的定义并产生满足 Zipf 分布的不确定图数据, 对确定算法进行不确定化使之满足应用要求。实验结果表明, 不论是在确定图数据、不确定图数据还是人工数据集、真实数据集上, 改进的 LinLogLayout 算法都具有较好的聚类效果。实验结果也表明, 不确定性的存在对聚类结果具有不可忽略的影响。

关键词: 不确定图; 图挖掘; 能量模型; 模块化聚类; 图聚类

中图分类号: TP391.41

文献标志码: A

文章编号: 1001-3695(2012)08-3173-03

doi:10.3969/j.issn.1001-3695.2012.08.099

Clustering uncertain graphs through energy function and modularity optimization

DING Yue¹, ZHANG Yang^{1,2}, WANG Yong³, LI Wei-wei¹

(1. College of Information Engineering, Northwest A&F University, Yangling Shannxi 712100, China; 2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China; 3. School of Computer, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: In order to indicate that the presence of uncertainty has a clustering effect can not be ignored, this paper improved a algorithm called LinLogLayout which optimized LinLog and related energy models to compute layouts, and Newman and Girvan's Modularity to compute clusterings and enabled it to deal with uncertain graphs. In addition, it proposed an explicit definition of uncertain graph and generated uncertain graphs subject to Zipf distribution, and then related improvements made to the algorithm in order to meet the requirements. After evaluation on both certain graphs and uncertain graphs, synthetic datasets and real datasets, it demonstrates that the improved LinLogLayout algorithm can handle both certain and uncertain graphs well, meanwhile the results indicate that the presence of uncertainty has a clustering effect can not be ignored.

Key words: uncertain graph; graph mining; energy models; modularity clustering; graph clustering

作为最主流的数据模型之一,图数据结构广泛应用于社交网络建模、化合物分析、网络路由设计等方面。表1列举了一些图数据的应用领域和建模方式。然而大多数与网络(本文不特别区分网络与图、加权图与不确定图的区别,且所讨论的图数据均为无向图)相关的文献仅仅考虑边存在或者不存在,如定义1。同时,随着人们对于数据不确定性存在的认识逐渐提高,不确定性已经成为评价知识的重要指标^[1]。不确定图数据如图1所示,不确定性表现在图的边以一定的概率存在,如定义2,甚至顶点也能够以特定的概率出现。在文献[2]中PPI(protein-protein interaction)网络被看做一类不确定图,其顶点表示蛋白质,边表示蛋白质交互,边以不同的概率存在。在文献[3]中,无线网络被看做不确定图,不确定路由链接计算机终端或服务器构成无线网络。如果系统建构成图结构,系统的元素是节点,交互是边,每一个子系统相当于一个社区。两个最具代表性表述网络结构的方法是布局 and 聚类,前者在一个空间中为节点分配位置,后者分割节点集合为单独的子集。文献[4,5]统一了 Girvan 和 Newman 的模块化理论,提出了一个聚类质量衡量标准,包括节点对两两之间的引力和斥力能量模

型和评价布局质量的标准。

由于不确定图数据无处不在以及不确定性的重要实践意义,本文改进并扩展了基于 Barnes-Hut 能量布局模型和 Girvan-Newman 模块化聚类的 LinLogLayout 算法^[4],以使其能够聚类不确定图数据。LinLogLayout 结合了计算图布局(定位图节点)和图聚类,可用于识别具有紧密联系的节点簇(像在社交网络中探寻朋友群或同事群,在 Web 网络中搜索相关文档,软件系统中识别关联子系统)。

表1 不同应用领域中实体的图形表示

应用	图形	顶点	边
生物信息学	蛋白质结构	氨基酸	接触残基
社交网络	社会关系	个体或组织	依赖关系
Web 挖掘	Web 浏览模式	Web 页面	页面间超链接
网络计算	计算机网络	计算机	机器间互联

1 能量布局模型和模块化聚类

首先给出确定图数据及不确定图数据的定义。

定义1 确定图。确定图 G 是一个四元组, $G = ((V, E),$

收稿日期: 2011-12-20; **修回日期:** 2012-02-01 **基金项目:** 国家自然科学基金资助项目(60873196); 中央高校基本科研业务费专项资金资助项目(QN2009092)

作者简介: 丁悦(1987-),女,硕士,主要研究方向为图数据挖掘(dyhlpt@163.com); 张阳(1975-),男,教授,博导,博士,主要研究方向为数据挖掘与机器学习; 王勇(1973-),男,副教授,硕导,博士,主要研究方向为数据挖掘与机器学习; 李伟卫(1988-),男,硕士,主要研究方向为分布式数据挖掘。

Σ_V, Σ_E, L 。其中: (V, E) 是一个无向图, V 是图 G 的顶点集合, $E \subseteq V \times V$ 是图 G 的边集合; Σ_V 和 Σ_E 分别是图 G 的节点标号集合和边标号集合; $L: V \rightarrow \Sigma_V, E \rightarrow \Sigma_E$ 是一个函数, 用来对顶点和边分配标号。

定义 2 不确定图。不确定图 UG 是一个五元组 $UG = ((V, E), \Sigma_V, \Sigma_E, L, P)$ 。其中: (V, E) 是一个无向图, V 是图 G 的顶点集合, $E \subseteq V \times V$ 是图 G 的边集合; Σ_V 和 Σ_E 分别是图 G 的节点标号集合和边标号集合; $L: V \rightarrow \Sigma_V, E \rightarrow \Sigma_E$ 是一个函数, 用来对顶点和边分配标号; $P: E \rightarrow (0, 1]$ 是边的存在可能性函数。边的存在可能性为 1 表示边一定存在。因此, 确定图(定义 1)可以看做所有边的存在可能性皆为 1 的特殊的不确定图。图 1 是一个不确定图模型。

1.1 (a, r) 能量布局模型

图数据的 d 维布局 p 映射每一个节点 v 到 R^d 空间一个具体的位置 p_v 。能量模型是衡量布局质量的核心。由于力是能量的负梯度, 能量模型被看做一个力系统, 能量越小意味着布局越好, 能量最小相当于力平衡^[6]。力的强度和欧氏距离的幂成比例。正式地, 给一个布局 p 和两个节点 $u, v, u \neq v, v$ 对 u 的引力定义为

$$w\{u, v\} \|p_u - p_v\|^a \overrightarrow{p_u p_v} \quad (1)$$

v 对 u 的斥力定义为

$$w_u w_v \|p_u - p_v\|^r \overrightarrow{p_v p_u} \quad (2)$$

其中: w_u 是节点 u 的权重, 每一个无序的节点对 $\{u, v\}$ (包含 $u = v$) 有一个非负的边权 $w\{u, v\}$; $\|p_u - p_v\|$ 是 u 和 v 在布局 p 下的欧氏距离; $\overrightarrow{p_u p_v}$ 是 u 指向 v 的单元长度向量; a 和 r 是实数常量, 且 $a > r$, 确保连接的节点之间的引力增长速度大于斥力, 从而阻止孤立节点的无限距离。在最广泛使用的 LinLog 能量模型^[4]中, $a = 0, r = -1$ 。

给定一个布局 p 和常量 $a, r \in R$ 且 $a > r$, (a, r) 能量定义为

$$\sum_{\{u, v\}: u \neq v} (w_{\{u, v\}} \frac{\|p_u - p_v\|^{a+1}}{a+1} - w_u w_v \frac{\|p_u - p_v\|^{r+1}}{r+1}) \quad (3)$$

其中 $\frac{\|p_u - p_v\|^{-1+1}}{-1+1}$ 当做 $\ln \|p_u - p_v\|$ 处理。

最小化能量布局模型的工作流程如下: a) 随机分配初始布局给每一个节点; b) 每一次迭代通过计算能量函数的一阶导数为每一个节点的移动确定方向和距离, 尝试通过能量的变换改进布局。Barnes-Hut 能量模型算法^[7] 每次迭代的时间复杂度是 $O(|E| + |V| \log |V|)$, 能够满足本文的要求。

1.2 模块化聚类测量

模块度确定一个网络划分社区的质量。一个好的划分, 其模块度很高, 即社区内部的联系特别紧密, 社区之间的联系特别稀疏。社区检测和聚类两者并不完全一样, 不像传统的层析聚类方法试图说明哪些边对于社区是最主要的, 它关注那些次要的常处于两个社区之间的边。Girvan-Newman 算法是典型的在复杂系统中检测社区的方法。它的最终结果是一个系统树图, 随着 Girvan-Newman 算法的运行, 系统树图从上至下建立 (逐步地移除链接分裂网络为不同社区)。

考虑节点权重是任意的, 一个聚类 p 的模块度定义为

$$\sum_{c \in p(V)} \left(\frac{w\{c, c\}}{w\{V, V\}} - \frac{\frac{1}{2} w_c^2}{\frac{1}{2} w_V^2} \right) \quad (4)$$

其中: V 是图中所有节点的集合; $p(V)$ 是簇的集合, 权值函数很自然地延伸到节点集合和边集合; $w\{c, c\}$ 为簇 c 中所有边的权值的总和; w_c 是簇 c 中所有节点的权值总和。

布局和聚类两者相互补充。一方面, 聚类能够揭示易于理解的高维的结构; 另一方面, 只有布局能够揭示簇之间的关系、簇的内部结构及节点与簇的关系。基于多层次的模块化聚类算法如下:

算法: 多层次聚类。

输入: 图 (graph), 粗化器 (coarsener), 精炼器 (refiner), 缩小因子 (reduction factor);

输出: 聚类结果。

//粗化过程 level[1] $\leftarrow$ graph;

for l from 1 to $\dots$ do

level[l+1] $\leftarrow$ coarsener (level[l], reduction factor);

if no clusters merged then break;

//精炼过程

clustering $\leftarrow$ vertices of level[l_{max}];

for l from l_{max} - 1 to 1 do

project clustering from level[l+1] to level[l];

clustering $\leftarrow$ refiner (level[l], clustering);

从单个簇开始执行一个贪婪的粗化过程, 选择能够导致最大模块度增加量的簇合并。粗化过程后的中间结果无论何时, 簇的数量以一个确定的百分数减少都被记录为一个粗化层次, 缩小因子起决定性作用。每一个粗化层次都是一个图, 图的节点是代表各自粗化状态的簇。多次粗化过程后进入精炼过程, 精炼过程被应用到每一个粗化层次, 不断重复地在簇之间移动节点。本文使用文献[8]中提出的多层次算法, 实验证明它能比传统的单层次算法产生更好的聚类结果。

2 实验

2.1 数据集

分别在真实数据集和人工数据集 (不确定性人工产生) 上进行实验。从 Pajek 数据集获取两个边的权重默认为 1.0 的确定数据集 Hitech 和 USAir97。为满足对比实验要求, 基于 Zipf^[9] 分布产生一组相同数量的浮点数作为边的权重, 使每条边具有在 $(0, 1]$ 内的权重表示边的存在概率, 从而生成相应的不确定图数据集。

此外还获取了新浪微博和人人网数据集。笔者的数据挖掘小组成功开发的“何时我最忙”微博第三方应用使得本文获取新浪微博的数据成为可能。其获取方法如下: 使用 OAuth (开放授权), 基于 SDK 使用脚本语言 PHP 获取数据集。本文仅仅考虑普通用户的好友关系, 并认为 A 请求 B 为好友与 B 请求 A 为好友的概率相同。人人网数据集的获取方法如下: 首先, 有一个页面引导用户获得认证, 若认证成功, 服务器会返回一个 session_key, 应用可以通过调用此 session_key 请求数据。两者的不同为: 人人网反映了同学或朋友之间相对真实可靠的好友关系 (相互准许方可加为好友), 然而大多数参与者参与新浪微博仅仅因为偶像效应, 因此好友关系不是非常可靠, 甚至无意义。

实验中所用到的数据集形式如表 2 所示。第一列是数据集的名称, 第二列是边的个数, 第三、四列分别是数据集的类型和节点的意义, 最后一列给出了数据集的表现形式。

表2 实验中所用到的数据集

数据集	边数	类型	节点	数据形式
新浪微博	3346	社交网络	用户 ID	1627538970 1627543431 1
人人网	794	社交网络	用户 ID	42051169 51923155 0
Hitech	147	工作网络	雇员	Bob Ovid 0.3
USAir97	2126	交通网络	机场	Chicago SanFrancisco 0.1

2.2 实验设置

整个程序包含计算图布局 and 图聚类两个主要任务。从一个文件中读取图数据,计算布局并进行聚类,并将结果写入另一个文件,同时以对话框的形式展示聚类结果。由于初始布局是随机的,为了得到更好的布局,理论上,应尽可能多地增加迭代的次数,特别当最后 5% 的迭代能量依旧在不断地降低的情况。迭代次数和图的规模应成比例。考虑实际使用的数据集的规模,本实验设置迭代次数为 100;表示斥力距离的幂 repuExponent 为 0.0 (即对数斥力);表示引力距离的幂 attrExponent 为 2.0;重力因子 gravFactor 为 0.2,重力吸引所有节点到布局的重心,可以阻止未连接节点飘得太远甚至扭曲布局。程序采用 Java 实现,基于 AMD Turion 1.79 GHz CPU 及 2.0 GB 内存的计算机。

2.3 实验结果分析

从实验结果可以看出,图的规模越小,聚类效果越好。布局(节点位置)和聚类(节点颜色)能够很好地归类确定数据集和不确定数据集的不同规模的簇。图2表明了一个科技公司中两个雇员之间的好友关系,不同的颜色反映出一定范围内关系密切的好友关系圈。从图3中可以容易地看出不同机场直达航班及非直达航班的情况。图4揭示了真实不确定数据集的聚类结果。微博数据集揭示出使用笔者应用的参与者之间的关系,人人网数据集揭示了一个特定用户的好友关系。可以看出,实验结果能够比较客观地表现实际中的群聚关系,同时验证了不确定性的存在对聚类结果具有不可忽略的影响。

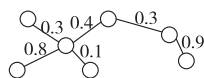


图1 不确定图模型

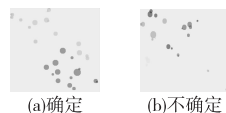


图2 Hitech 数据集的实验结果

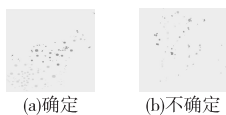


图3 USAir97数据集的实验结果

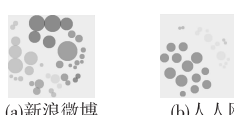


图4 真实不确定数据集的实验结果

3 结束语

本文在不同类型、不同规模的数据集上测试改进的 LinLogLayout 算法,这对于人们认识不确定性的重要性有很大的意义。事实上,对于小规模的不确定图数据,不管是收集还是分析,对研究者和应用设计者来说都是相对容易的,然而之前并没有展开针对不确定数据集的实验。未来将考虑如何将 LinLogLayout 算法进行改进,从而可以以相对较低的时间复杂度处理大规模图数据。此外,随着越来越多的图数据以数据流的形式出现,笔者已经着手设计新的算法使之可以处理图数据流聚类,甚至是不确定图数据流。

参考文献:

- [1] 周傲英,金澈清,王国仁,等. 不确定性数据管理技术研究综述[J]. 计算机学报,2009,32(1):1-16.
- [2] ASTHANA S, KING O. D, GIBBONS F D, *et al.* Predicting protein complex membership using orobabilistic network reliability[J]. *Genome Research*,2004,14(6):1170-1175.
- [3] GHOSH J, NGO H Q, YOON S, *et al.* On a routing problem within probabilistic graphs and its application to intermittently connected networks[C]//Proc of IEEE Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies. Washington DC:IEEE Computer Society,2007:1721-1729.
- [4] NOACK A. Energy models for graph clustering[J]. *Journal of Graph Algorithms and Applications*,2007,11(2):453-480.
- [5] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. *Physical Review E*,2004,69(2):56-68.
- [6] BRANDES U. Drawing graphs: methods and models [M]. London: Springer-Verlag,2001:71-86.
- [7] BARNES J, HUT P. A hierarchical $O(N \log N)$ force-calculation algorithm[J]. *Nature*,1986,324(4):446-449.
- [8] NOACK A, ROTTA R. Multi-level algorithms for modularity clustering [C]//Proc of the 8th International Symposium on Experimental Algorithms. Berlin: Springer-Verlag,2009:257-268.
- [9] KALI R. The city as a giant component: a random graph approach to Zipf's law[J]. *Applied Economics Letters*,2007,10(11):717-720.