基于 Pagerank 的社会网络关键节点发现算法

仇丽青，陈卓艳

（山东科技大学 信息科学与工程学院，山东 青岛 266590）

摘　要：受 Pagerank 算法启发，将社会网络中的节点模拟成 Web 中的页面，将边模拟成 Web 中的超链接，提出基于 Pagerank 的社会网络关键节点发现算法。通过实验验证了该算法的可行性。

关键词：Pagerank；社会网络；关键节点；挖掘分析

中图分类号：TP312　　**文献标识码**：A　　**文章编号**：1672-7800(2014)008-0048-02

0　引言

作为复杂网络的一个组成部分，社会网络是一个抽象的社会结构，其实质上是人与人之间复杂的关系网络。社会网络可以看作是由相互连接的节点和边构成，其中节点是现实中参与社会活动的个人或组织等，边是指活动参与者之间的关系或联系。常见的社会网络包括合作者网络、电子邮件网络和互联网社区等。随着 Internet 的快速发展和各种社交网站的出现，许多大型社会网络的数据可以从运营商或互联网上获得。因此，人们开始把社会网络引入到数据挖掘领域，使之成为一个新的研究方向。

对社会网络的挖掘分析，一个非常重要的问题是关键节点进行发现。对关键节点进行挖掘有着广泛的应用，例如在市场营销、计算机网络安全防御等方面。在市场营销方面，可以通过关键节点挖掘识别出具有影响力的潜在客户，对这些客户进行新产品的免费试用，从而使其对周围亲戚朋友起到积极的影响作用，带动他们也来购买产品，他们又会影响他们自己的朋友，口口相传，实现“病毒式营销”。在计算机网络安全防御方面，可以先挖掘出那些容易传播病毒的关键计算机，集中精力对他们进行防御，而不用盲目对整个网络进行防御，从而达到降低成本提高效率的目的。

在社会网络中，挖掘最具影响力节点的最优问题是一个 NP 问题[1]。目前往往采用贪婪算法来解决该问题，虽然所选节点的影响力可以得到保证，但是当挖掘大型网络时，效率低下，算法运行时间难以接受。针对这一问题，本文提出了一种新算法——基于 PageRank 的社会网络关键节点发现算法。该算法思想来源于互联网中著名的 PageRank 算法[2-3]，并对其进行改进，使之适合于社会网络。实验结果表明，该算法不仅能够保证挖掘的影响度，而且时间效率显著提高。

1　相关研究

近年来，社会网络的关键节点挖掘在研究上获得了广泛关注。关于社会网络上关键节点的影响力，学术界没有一个统一的标准。一般认为，节点的影响力可以用一种打分函数进行衡量，而这个打分函数的取值可以被理解为节点在社会网络上的影响力，函数值大的节点被认为是社会网络中的关键节点。

中心度分析是社会网络中关键节点挖掘的一个重要方法，比较经典的中心度计算方法包括：度中心度、紧密中心度、间距中心度等。这些中心度分析方法从不同侧面衡量节点的重要性：度中心度是用节点度数来衡量的中心度，紧密中心度是依据网络中各节点之间的紧密性或距离而测量的中心度，间距中心度则是节点与其它节点之间相间隔的程度。这些中心度分析方法的时间复杂度和空间复杂度往往较高[4]。例如，对于一个包含 n 个节点和 m 条边的社会网络，紧密中心度的时间复杂度为 $O(n^2)$，间距中心度的时间复杂度为 $O(nm)$。这对于大型复杂社会网络来说显然不适用。另外，Kempe 等人已经证实了关键节点的挖掘是一个 NP 问题，并提出采用贪婪算法来解决这个问题，但是贪婪算法最大的问题在于它的时间复杂度很高。在这种情况下，本文提出采用 Pagerank 算法进行关键节点挖掘，实验证明取得了较好的结果。

2　算法提出

2.1　问题定义

给定一个无向无权图 $G(V,E)$，挖掘 k 个关键节点，使得从这 k 个关键节点出发的消息传播最大化。其中，消

基金项目：国家博士后项目(2013M541938)；山东省博士后创新项目(201302036)

作者简介：仇丽青(1978－)，女，山东德州人，博士，山东科技大学信息科学与工程学院讲师，研究方向为数据挖掘、社会网络。

息传播模型的限制条件为：如果节点 i 能够影响或激活节点 j，那么节点 i 在距离 d 内必须可达节点 j。换言之，需要计算在给定节点集合 B 时，能够被集合 B 影响到的节点数量 $\delta(B)$ 最大：

$$\max_{B\subseteq V}\delta(B) s.t. \mid B \mid \leqslant k \tag{1}$$

2.2 Pagerank 算法

Pagerank 算法由拉里·佩奇和谢尔盖·布林于 1998 年提出并发表，该算法是一个非常经典的 Web 页面排名算法。凭借着 Pagerank 算法的基本思想与成功的商业经营，Google 已经成为全球非常优秀的互联网企业。Pagerank 算法的思想主要是基于网络结构的链接分析。它基于这样一种基本思想：被用户访问越多的网页其质量越高，而用户在浏览网页时主要通过超链接进行页面跳转，因此可以通过分析超链接组成的拓扑结构来推算每个网页被访问频率的高低。假设当一个用户停留在某个页面时，跳转到页面上每个被链接页面的概率相同。对于页面 p_i，它的 Pagerank 值定义为：

$$Pagerank(p_i) = \frac{1-q}{N} + q\sum_{p_j} \frac{Pagerank(p_j)}{L(p_j)} \tag{2}$$

其中，q 是阻尼系数，一般定义为 0.85；$p_1, p_2, \cdots, p_N$ 是被研究的页面；$L(p_j)$ 是 p_j 链接页面的数量；N 是所有页面的数量。

2.3 基于 Pagerank 的社会网络关键节点发现算法

Pagerank 算法是针对于 Web 系统而设计，其基本思想是，指向某页面的链接将增加该页面的 Pagerank 值。受该算法启发，可以将社会网络中的节点模拟成 Web 中的页面，而将社会网络中的边模拟成 Web 中的超链接。因此，可以采用类似 Pagerank 算法来计算节点的影响力。

为了简便起见，仅考虑社会网络中最简单的无向无权图，算法如下：

Algorithm: Pagerank
initialize $B = \phi$
set $Pagerank(v) = 1$ for all $v \in V$;
for $i = 1$ to k do
　while ($Surplus > \varepsilon$){
　　for each neighbore $N_v \in V \backslash B$ do
　　$PR(N_v) = 0$
　　end for
　　while ($N_v \neq \phi$){
　　for($j = 1 \cdots \mid N_v \mid$)
　　$Pagerank(N_v) += Pageranke(N_v)/d_{N_v}$
　　end for
　end while
　$Pagerank(N_v) = (1-d)/N + d * Pagerank(N_v)$
　$Surplus = Pagerank(v) - Pagerank(N_v)$
　$Pagerank(v) = Pagerank(N_v)$
　$S = S \bigcup \{\text{argmax}_{v\in V\backslash B}\{Pagerank(v)\}\}$
　end while
end for
output B.

该算法的输入为网络 $G(V,E)$ 和常数 k，输出为节点集合 B。首先把该网络中所有节点的 Pagerank 值设定为 1，然后执行 k 次循环，在每次循环中通过计算节点的 Pagerank 值选出最大的节点加入到集合 B 中。整个算法是迭代进行的，直至得到一个包含 k 个节点的集合 B。

3 实验验证

本文采用 Karate 俱乐部作为实验数据。该网络是社会网络分析领域的经典数据集。20 世纪 70 年代初期，社会学家 Zachary 用了两年时间观察美国一所大学空手道俱乐部 34 名成员间的社会关系。基于这些成员在俱乐部内部及外部的交往情况，他构造了成员之间的社会关系网。该网络包含了 34 个节点，两个节点之间有一条边则意味着相应的两个成员之间至少是交往频繁的朋友关系。Karate 俱乐部社会网络如图 1 所示。

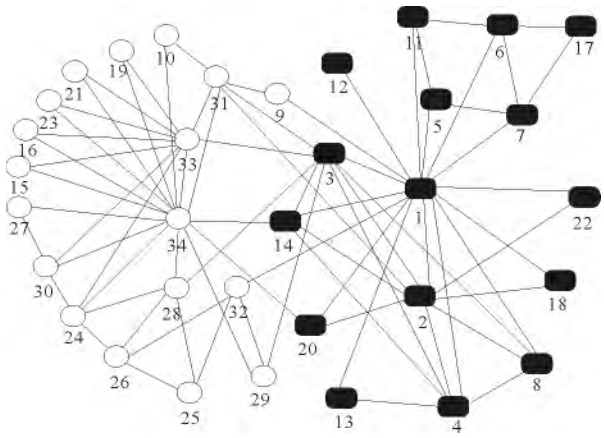

图 1　Karate 俱乐部

采用本文所提出的 Pagerank 算法对这 34 个节点进行排序，选出 Pagerank 值排名前 10 位的节点如表 1 所示。

表 1　Pagerank 排名

排名	Node ID	Pagerank 值
1	34	0.101
2	1	0.097
3	33	0.072
4	3	0.057
5	2	0.053
6	32	0.037
7	4	0.036
8	24	0.032
9	9	0.030
10	14	0.030

经过比较发现，上述结果符合实际排名。例如节点 34 和节点 1 分别是该俱乐部中实际的两个核心节点。

4 结语

社会网络关键节点发现是社会网络分析的一个重要

基于局部克里金插值算法的横断层三维建模

林永良

(天津城建大学 信息化建设管理中心,天津 300384)

摘 要:分析了现有断层建模技术,提出了基于局部克里金插值算法的横断层自动分析和显示技术。算法通过搜索遍历钻孔点计算模型的有效区域和阈值,选用基于面模型的网格三角化法构建地质体,结合 OpenGL 实现了三维构建和漫游。实验证明,该方法能够实现横断层的自动分析和显示。

关键词:横断层;三维建模;Kriging 插值算法;三维可视化

中图分类号:TP312 **文献标识码**:A **文章编号**:1672-7800(2014)008-0050-03

0 引言

简单连续地质体的三维建模技术较为成熟,有基于面模型的 DEM、TIN 建模技术,基于体模型的 TEN、Octree 建模技术,以及基于混合模型的 TI-CSG 和 TIN-Octree 等。而针对破坏地质连续性的断层结构研究则较为缺乏,其主要原因是断层迫使连续插值计算方法不能满足需求。本文提出基于局部克里金(Kriging)插值算法来实现简单断层分析和模拟,通过分析原始地质数据的相关特性,设置计算精度,进而将地质体自动划分成不同区域,最后针对不同区域分别进行 Kriging 插值和三维建模。

1 三维地质建模相关技术

1.1 三维地质建模技术

本文根据面模型、体模型和混合模型的优缺点,考虑到数据的存储结构及可视化问题,最终选择了规则网格的三角化建模方法来构建地质层。该方法的数据结构简单、存储量小,便于数据的存取和检索,提高了差值运算的效率,缩短了模型的渲染时间。不仅如此,运用这种方法还有利于实现模型之间的转换,即如果将相邻层面间对应的三角形顶点相连,则能够形成基于类三棱柱的体模型建模。

1.2 断层模型构建方法

三维断层建模方法主要有整体法、局部法和统一建模法[1]。整体法是基于整体地层恢复的断层构建方法,它适用于断距小、断层两侧地层的形态及厚度存在一定相似性的断层系统。局部法是基于分区插值的断层构模技术,适用于同沉积断层和部分大型走滑、倾滑断层,能够反映断层的准确位置,结构简单、运算速度快,符合本文的需求。统一构模技术是基于上述两种方法的结合,它既适合于同沉积断层,也适用于非同沉积地层系统,还可应用于断层终止于地层内部的情况,但由于它所需考虑因素多,结构复杂,因而导致其存储容量大、构模方式复杂多变。

1.3 三维空间插值算法

用于地质体建模的数据(钻孔数据)具有分散性和不规则性,必须对其进行数据插值才能得到较为完整的地质数据。常用的插值法有最小二乘距离加权插值法、线性内插法和Kriging插值法等[2]。前两种方法的误差较大、精

研究方向。受 Pagerank 算法启发,本文提出了基于 Pagerank 的社会网络关键节点发现算法。将节点模拟成 Web 中的页面,将链接模拟成 Web 中的超链接。实验证明,将 Pagerank 算法应用于社会网络中的关键节点发现取得了较好的结果。

参考文献:

[1] KEMPE D, KLEINBERG J M, TARDOS E. Maximizing the spread of influence through a social network[C]. The 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2003:137-146.

[2] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[D]. Stanford :Stanford University, 1998.

[3] BIANCHINI M, GORI M, SCARSELLI F. Inside pagerank[J]. ACM Transactions on Internet Technology, 2006(2):92-128.

[4] WASSERMAN S, FAUST K. Social network analysis: methods and applications[R]. 1994:3-5.

(责任编辑:孙 娟)

作者简介:林永良(1986-),男,陕西咸阳人,硕士,天津城建大学信息化建设管理中心实验师,研究方向为计算机应用技术、虚拟现实、智能信息处理。