

基于免疫克隆算法和 K-medoids 的网络节点聚类

李 永^{1,2}, 余镇危¹

(1 中国矿业大学(北京) 机电与信息工程学院, 北京 100083; 2 盐城师范学院 信息科学与技术学院, 江苏 盐城 224002)

摘 要: 在大规模分布式网络应用中, 对网络节点进行聚类是构建高效网络体系结构的有效办法之一. 在利用网络坐标系统 Vivaldi 得到各个节点的网络坐标的基础上, 对网络节点进行 K-medoids 聚类. 然后, 针对 K-medoids 算法对初始中心选值敏感和易陷入局部极值的问题, 提出基于免疫克隆算法的 K-medoids 聚类. 实验结果表明, 该聚类算法具有良好的可靠性及可扩展性, 能对节点进行有效聚类.

关键词: 网络节点聚类; 免疫克隆算法; K-medoids; Vivaldi

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)08-0119-04

Network Nodes Clustering Based on Immune Clonal Algorithm and K-medoids

LI Yong^{1,2}, YU Zhen-wei¹

(1 College of Mechanical and Electronic Engineering, China University of Mining and Technology, Beijing 100083, China;
2 School of Information Science and Technology, Yancheng Teachers College, Yancheng 224002, China)

Abstract: In large-scale distributed network applications, nodes clustering is a useful way to construct an effective network infrastructure. The coordinates of network nodes can be get by the network coordinates system Vivaldi, then, network nodes can be clustered by the K-medoids algorithm according to their coordinates. But K-medoids is sensitive to the initial cluster centers and easy to get stuck at the local optimal solutions. In order to improve the performance of the K-medoids algorithm, the K-medoids based on immune clonal algorithm(KICA) is presented in this paper. Experimental results show KICA has good reliability and expansibility, and it is effective for clustering internet nodes.

Key words: network nodes clustering; immune clonal algorithm; K-medoids; Vivaldi

1 引言

随着 Internet 在全球范围的飞速发展, 应用层组播、P2P 网络、大规模文件存取、文件多点下载等大规模分布式的网络应用在 Internet 服务中占的比重越来越大. 在这些网络应用中, 如何根据网络节点的分布状况构建一个高效的网络体系结构是非常重要的^[1]. 一种行之有效的办法是利用网络节点之间的距离对网络节点进行聚类, 将地理上邻近的节点分到同一个聚簇中, 同一个簇中的节点网络距离较近, 不在同一个簇中的节点网络距离较远, 从而达到

减少大规模分布式网络应用中网络流量, 提高路由效率的目的.

为了获知节点之间在地理上的关系, 将地理上邻近的节点分到同一个聚集中. 文献[2]假定每个自治系统都有一个 SNMP 网管, 通过该网管可以获得自治域内的静态拓扑结构. 这种方法需要自治系统内所有的路由器、交换机以及端系统都支持 SNMP 协议, 目前显然不太实用. 文献[3]直接从 BGP 路由器上获取拓扑信息, 节点根据它们的 IP 网络地址被分组成簇. 但由于 Internet 网络中端节点没有 BGP 路由表的访问权限, 该方法对普通端节点根本无法

收稿日期: 2011-05-29; 修回日期: 2011-06-29

基金项目: 中央高校基本科研业务费专项资金(2011YJ)

适用. 文献[4]提出了一种基于静态路标的节点聚集算法, 但该方法必须事先指定若干静态路标作为测量的基点, 当端节点数目很多的时候, 将使得充当静态路标的节点成为热点, 面临着诸如集中式地标带来的单点失效以及非线性计算开销较大等问题. 文献[5]通过节点的网络坐标并计算出与其他节点之间的网络距离, 根据网络距离使用 K-means 算法进行网络节点聚类, 但由于随机选取聚类数和初始聚类中心, 对聚类结果影响较大.

本文采用网络坐标系 Vivaldi^[6] 将网络节点映射到坐标空间中, Vivaldi 是一种简单的、轻量级的网络坐标系, 聚类算法以对坐标系中的点进行聚类的形式对网络节点进行聚类. 考虑到网络节点为离散的实际点, 采用对孤立点具有很好的健壮性的 K-medoids 算法, 并针对 K-medoids 算法易陷入局部极值的问题, 结合免疫克隆算法的高效率全局优化能力提出 KICA (K-medoids based on immune clonal algorithm) 算法. 此算法在进行网络节点的聚类时, 一方面能够限制不良个体生成, 缩小搜索空间, 防止较优个体在进化中的扩散; 另一方面, 能够加快收敛速度, 提高算法在解空间的寻优能力.

2 网络节点的 K-medoids 聚类

2.1 K-medoids 简介

目前的聚类方法很多, 其中基于划分的聚类算法主要包括 K-means 算法和 K-medoids 算法^[7]. 这两种算法很相似, 都采用某个空间距离作为评价相似程度的指标; 然后确定某个评价聚类划分结果质量的目标函数, 在给出初始聚类中心点后, 用迭代法找出使目标函数取极值的最好聚类划分结果. 其缺点是受初始值影响大, 通常只能得到局部最优解. 这两种算法的区别是 K-means 算法采用节点均值(质心)作为聚类中心, 是一个抽象点, 不是具体的节点; 此外, K-means 算法对孤立点非常敏感. 而 K-medoids 算法采用实际节点作为聚类中心, 可以很好的解决 K-means 算法对孤立点非常敏感的问题, 大大的提高聚类的精度.

2.2 网络节点的 K-medoids 聚类算法

在利用网络坐标系 Vivaldi 得到各个节点的网络坐标的基础上, 对网络节点进行 K-medoids 聚类. 在一个具有 n 个节点的网络中, 每个网络节点 i 都有一个来自 Vivaldi 的 m 维网络坐标向量 $v_i = \{v_{i1}, v_{i2}, \dots, v_{im}\}$, $i = \{1, 2, \dots, n\}$, 表示网络节点映射为 m 维网络坐标空间 R^m 中的一个点. 因此, 网络

节点的样本空间为 $V = \{v_1, v_2, \dots, v_n\}$ 可表示为

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} = \{v_{ij}\}_{n \times m} \quad (1)$$

式中, v_{ij} 为节点 i 的第 j 维的坐标值, $i \in \{1, 2, \dots, n\}$, $j \in \{1, 2, \dots, m\}$. 聚类算法就是将这 n 个 m 维的节点坐标空间点分成 K 个簇 $V_1, V_2, \dots, V_K$, 同时令 $C = \{V_1, V_2, \dots, V_K\}$, 这 K 个簇的中心点的坐标向量分别为 $c_1, c_2, \dots, c_K$, 令 $c = \{c_1, c_2, \dots, c_K\}$. 同一簇内的节点具有较高的相似度, 不同簇内的节点具有较低的相似度, 相似度的高低利用网络节点间的距离进行度量. 对于 $\forall i \neq j, V_i \cap V_j = \Phi$, 且 $V = \bigcup_{1 \leq k \leq K} V_k$. $|V|$ 表示整个样本空间的节点数即网络节点数 n , $|V_k|$ 表示第 k 个聚簇内的网络节点数. 在 K-medoids 聚类算法中, 节点间的相似度高, 即各节点到对应簇中心的距离(目标函数)小. 用 $d(v_i, c_k)$ 表示 v_i 到 c_k 的 K-medoids 聚类距离算子, 代表 v_i 到 c_k 的距离成本, 在本文中用欧氏距离来表示, 如式(2)所示, 其中 v_i 和 c_k 分别是网络节点和簇中心的坐标向量.

$$d(v_i, c_k) = \|v_i - c_k\| = \sqrt{\sum_{j=1}^m (v_{ij} - c_{kj})^2} \quad (2)$$

为了反映网络节点 i 与簇 V_k 的隶属程度, 引入隶属度函数如下:

$$u_{ik} = \begin{cases} 1 & (d(v_i, c_k) = \min\{d(v_i, c_h)\}) \\ 0 & (d(v_i, c_k) \neq \min\{d(v_i, c_h)\}) \end{cases} \quad (3)$$

$1 \leq h \leq K, v_i \notin c$

因此, 由式(2)和式(3), 目标函数可定义为

$$J(c) = \sum_{k=1}^K \sum_{i=1}^n u_{ik} d(v_i, c_k) \quad (4)$$

综上, 网络节点聚类问题的数学模型如下:

$$\begin{aligned} & \text{Min } J(c) \\ & \text{s. t. } C = \{V_1, V_2, \dots, V_K\} \\ & c = \{c_1, c_2, \dots, c_K\} \\ & \forall i \neq j, V_i \cap V_j = \Phi \\ & V = \bigcup_{1 \leq k \leq K} V_k \end{aligned} \quad (5)$$

网络节点 K-medoids 聚类算法的流程如下:

(1) 从 n 个网络节点中, 随机选择 K 个节点作为簇中心.

(2) 根据式(3) 确定每个非簇中心节点的隶属度.

(3) 根据式(4) 计算目标函数 $J(c)$, 如果其值小于给定阈值, 或算法达到最大迭代次数, 停止运行并输出结果. 否则, 继续.

(4) 随机在簇内选择一个非簇中心的节点代替当前的簇中心, 转向第(2) 步.

由于 K-medoids 聚类算法采用的是基于目标函数的算法, 常用梯度法求解极值, 其搜索方向沿着能量减小的方向进行, 使得算法很容易陷入局部极值, 并且算法对初始中心的选择敏感. 因此, 需要对 K-medoids 聚类算法进行改进.

3 基于免疫克隆的 K-medoids 聚类算法

3.1 算法流程

算法流程如图 1 所示. 在本文的算法中, 把要聚类的网络节点样本空间中的对象看成抗原, 把每一次迭代产生的聚类中心看作免疫系统中的抗体, 并计算其亲和度, 进行克隆、变异等操作. 该算法的过程就是系统不断迭代, 产生抗体, 识别抗原, 再次迭代, 最后产生出可以捕获抗原的最佳抗体.

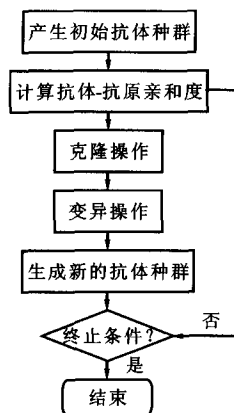


图 1 算法流程

3.2 算法详解

(1) 抗体编码与初始化抗体群

本算法采用实数编码, 一个抗体由 K 个簇中心组成. 抗体群的初始化方法是从网络节点坐标的样本空间 V 中随机选 K 个样本作为一个抗体体进行编码, 重复进行 N 次, 得到规模为 N 的抗体群 A .

(2) 抗体—抗原亲和度的确定

在免疫算法中, 抗体—抗原的亲和度常用群体中个体的适应度函数来表示. 本算法采用的亲和度函数为式(6).

$$F(s) = 1/J(s) \quad (6)$$

式中, $J(s)$ 代表所有网络节点到抗体 s 所表示的簇中心的最小欧氏距离和, $J(s)$ 可由式(4) 求出, $J(s)$ 越小, $F(s)$ 就越大, 表示抗体 s 越接近最优解. 利用

式(6) 计算抗体群 A 中各抗体的亲和度.

(3) 克隆操作

在一个种群里, 克隆的规模应是抗体—抗原的亲和度的单调递增函数. 对抗体群 A 中的每一个抗体 A_i 按 q_i 的规模克隆到临时抗体群 A^* 中, q_i 满足式(7).

$$q_i = \text{Int} \left\{ N \times \frac{F(A_i)}{\sum_{i=1}^N F(A_i)} \right\} \quad (7)$$

N 为抗体群 A 的规模; $\text{Int}()$ 表示取整操作.

(4) 变异操作

对克隆产生的临时抗体群 A^* 进行变异操作, 变异的概率一般为常数, 变异算子按照抗体的基因位进行, 每个抗体上的聚类中心按照变异概率随机变异. 若变异后抗体的亲和度大于变异前的亲和度, 则变异后的抗体取代变异前的抗体, 反之, 保持原抗体不变. 经变异后生成临时抗体群 B .

(5) 生成新的抗体群

对经过克隆和变异操作生成的临时抗体群 B , 选出亲和度最高且互不相同的 N 个抗体组成新的临时抗体群 B^* . 同时, 随机产生规模为 N_0 的抗体群 A_0 , 选出亲和度最高的 N_1 ($N_1 < N$) 个抗体组成抗体群 A_1 , 用 A_1 代替 B^* 中亲和度最低的 N_1 个抗体, 形成规模为 N 的新的抗体群 A .

(6) 终止条件

对于新生成的抗体群 A 中的抗体, 如果存在抗体, 即聚类中心, 满足最优条件, 则算法结束, 输出聚类结果. 否则, 转到(2) 继续操作.

4 仿真实验

采用通用的拓扑产生器 Brite 生成网络拓扑数据, 模拟底层的物理网络. 实验中生成包含 6 000 个节点的层次结构网络拓扑图, 分成 10 个自治域, 每个自治域包含 600 个节点, 自治域间和自治域内的节点之间延时分别控制在 40 ~ 100 ms 和 5 ~ 30 ms. 为了评价聚类算法的效果, 使用 $D(i, j)$ 表示节点 i 和 j 之间的延时, 使用平均簇中延时 ADC (average delay cluster) 作为评价指标, ADC 是网络中各个簇的簇中平均节点延时的平均值, 其定义如式(8).

$$\text{ADC} =$$

$$\frac{\sum_{k=1}^K \left(\sum_{i,j \in V_k \wedge i \neq j} D(i, j) / (|V_k| (|V_k| - 1)) \right)}{K} \quad (8)$$

网络中节点之间的平均延时 AND (average delay network) 定义如式(9). 这个值可以用来与

ADC 进行比较.

$$ADN = \frac{\sum_{i,j \in V \wedge i \neq j} D(i,j)}{n(n-1)} \quad (9)$$

为了评价 KICA 的聚类效果,每次选取不同的初始种群,进行了 24 次独立的 KICA 实验,当算法终止时记录下每次得到的最优聚类结果,并利用式(8)求出 ADC 的值. ADC 与 ANC 的比较如图 2 所示,聚类后平均簇中延时比聚类前网络节点的平均延时低很多,说明聚类算法是可靠的, KICA 对网络节点进行了有效的聚类.

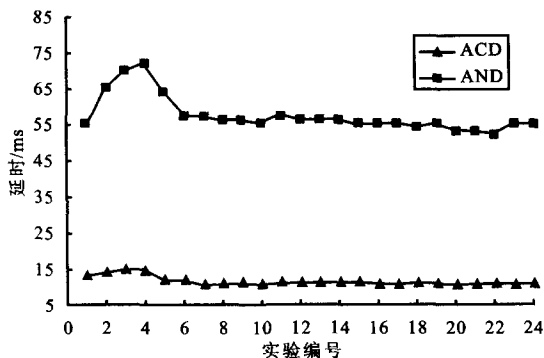


图 2 KICA 聚类效果

为了比较 KICA 与 K-medoids 以及 KGA (K-medoids based on genetic algorithm) 算法的性能,分别对它们进行了 24 次独立的实验,在每次实验中重新选取初始聚类中心和种群,相同编号的实验中 KICA 和 KGA 的初始种群相同. 从图 3 可以看出, KICA 聚类效果最好且性能稳定,选择不同初始种群时都能收敛到一个相对稳定的全局最优值. K-medoids 算法对初始聚类中心敏感,很容易陷入局部极值,当选择不同的初始聚类中心,算法终止时得到聚类的 ADC 值起伏较大. KGA 在全局搜索中的会出现早熟收敛问题,最终得到的聚类效果不如 KICA 好.

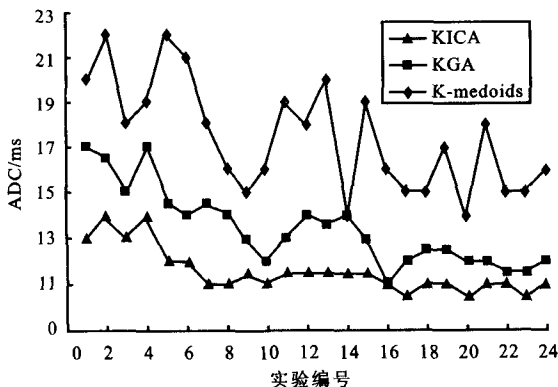


图 3 算法性能比较

5 结束语

本文提出一种基于网络节点坐标的聚类算法,采用网络坐标系统 Vivaldi 得到节点的坐标,然后对网络节点进行 K-medoids 聚类. 针对 K-medoids 算法对初始中心敏感、易陷入局部极值的问题,进一步提出网络节点的 KICA 聚类算法. 模拟实验结果表明利用 KICA 聚类算法的聚类效果好,算法有较高的可靠性. 另外,由于采用 Vivaldi 预测节点间的距离,当网络规模比较大时不会带来节点之间端到端的测量负担,算法具有较高的可扩展性. 今后的研究工作,一是降低网络坐标系统本身误差对节点聚类效果的影响. 二是在网络节点聚类的基础上研究 P2P、应用层组播等大规模分布式应用,实现减少网络流量,提高路由效率.

参考文献:

- [1] 陈阳,邓北星,李星. 基于坐标的网络节点聚类在 Internet 中的实验研究[J]. 大连理工大学学报,2005,45(增刊):41-43.
- [2] 黄道颖,黄建华,庄雷,等. 基于主动网络的分布式 P2P 网络模型[J]. 软件学报,2004,15(7):1081-1089.
- [3] Balachandrer Krishnamurthy, Jia Wang. On network-aware clustering of web clients [C]// Proceedings of ACM SIGCOMM 2000. Sweden: Stockholm, 2000.
- [4] Xin Yan Zhang, Qian Zhang, Zhengsheng Zhang, et al. A construction of locality-aware overlay network: overlay and its performance [J]. IEEE Journal on Selected Areas in Communications, 2004, 22(1):18-28.
- [5] Jiang Yi, Shi Hongbo, You Jinyuan. Network coordinates-based Peer-to-Peer hosts clustering [C]// Proc. of IFIP Int'l Conf. on Network and Parallel Computing. Tokyo, Japan, IEEE Press, 2006.
- [6] Dabek F, Cox R, Kaashoek F, et al. Vivaldi: a decentralized network coordinate system [C]// Proc. of ACM SIGCOMM'04. USA: Oregon, ACM Press, 2004.
- [7] 杨柳,钟诚,陆向艳. 基于 P2P 网络的分布式聚类算法研究分析[J]. 微电子学与计算机,2009,26(8):83-85.

作者简介:

李永男, (1978-), 博士研究生, 讲师. 研究方向为网络体系结构、分布式应用.

余镇危男, (1942-), 教授, 博士生导师. 研究方向为网络体系结构、覆盖网、下一代互联网.