基于时序特征的物联网社区划分算法①

王 杨, 黄亚坤, 张林静

(安徽师范大学 数学计算机科学学院, 芜湖 241000)

摘 要: 针对物联网环境下的社区划分问题, 提出一种基于时序特征的社区划分算法. 首先给出时序数据的相关定义; 然后对物联网社区的时序数据进行预处理, 提出了物联网社区划分算法, 并进行了算法性能分析. 通过实际网络社区数据的仿真实验表明该算法的高效性.

关键词: 时序特征; 区间模糊; 物联网; 社区; 划分算法

Community Partition Algorithm Based on Timing Characteristics Under Internet of Things

WANG Yang, HUANG Ya-Kun, ZHANG Lin-Jing

(School of Mathematics and Computer Science, Anhui Normal University, Hefei 230039, China)

Abstract: In order to solve service community dividing problems in the IOT environment, a new community partition algorithm which is based on timing characteristics on community was proposed. First, the paper gives the definition relevant to time series data; then presents timing data pre-processing schema. Finally, the proposed algorithm performance was analyzed. The simulation results show that the proposed algorithm is feasible.

Key words: timing characteristics; interval fuzzy; Internet of things; community; partitioning algorithm

随着物联网应用与研究的不断深入, 物联网环境中服务社区的构建与演化问题已经成为物联网服务提供机理研究的重要内容之一[1]. 物联网的服务模型、物联网服务需求的多维获取、物联网服务动态构建及动态演化技术等问题大多是基于物联网服务社区而展开.

在已有的复杂网络社区划分研究中[2], 社区划分算法可分为两大类: 第一类是基于图论的算法, 比如 Kernighan-Lin 算法[3]、基于 Laplace 矩阵谱平分法[4]、随机游走算法[5], 这类算法用矩阵表示网络中个体之间的关系, 然后计算矩阵最大特征值对应的特征向量, 通过 K-means 算法进行聚类划分; 第二类是层次聚类算法, 比如基于边介数度量的分裂算法[6]、GN 算法[7]、Newman 快速算法[8], 这类算法基本思想是不断从网络中移除边介数最大的边. 但文献[7]中的算法时间复杂性非常高, 文献[8]虽然将时间复杂性降至 $O(n^2)$,但划分的准确率较低, 因而均无法进行准确、高效的社区划分.本文提出的基于时序特征的物联网社区划分算法, 从时序特征考虑, 将物联网社区数据与时序数据相结合, 同时进行区间模糊处理, 基于边链接系数进行划分. 最后对该划分算法进行了仿真实验, 得出了该划分算法具有一定的可用性.

本文的主要工作是提出一种基于时序特征的物联网社区划分算法. 该方法首先对物联网社区数据进行时序特征处理, 然后根据时间序列进行社区的划分.

1 时序特征物联网社区预处理

1.1 时序数据的相关定义

定义 1. 时间序列: 时间序列是由记录值和记录时间组成的元素的有序集合, 记为:

$$X = \{x_1 = (v_1, t_1), x_1 = (v_2, t_2), \cdots x_n = (v_n, t_n)\},$$

元素 $x_i = (v_i, t_i)$ 表示时间序列在 t_i 时刻的记录值为 v_i, 记录时间 t_i 间是严格增加的[9]. 一般情况下, 某个时序数列采样间隔时间 Δt 相等, 可以看作 $t_1=0$ 作 $\Delta t=1$.

① 基金项目:中国博士后基金(20100480701);教育部人文社科青年基金(11YJC880119)
收稿时间:2012-03-19;收到修改稿时间:2012-05-14

此时将时间序列(时序数据):

$X=\{x_1=(v_1,t_1),x_1=(v_2,t_2),\cdots x_n=(v_n,t_n)\}$，简记为: $X=\{x_1,x_1,\cdots x_n\}$.

根据定义 1, 我们将社区节点的信息标号为连续的时序数据, 研究该时序数据的特征.

时序数据通常用一个实数来表达某一时刻的观测值, 这单一数值表示的信息并不十分充分; 在模糊集中, 一个时间点的隶属度依然是实数, 只是范围被限制在[0,1]中, 表达的元素更灵活、丰富. 下面给出区间模糊时序的定义:

定义 2. 区间模糊时序:

令 $d(i)(i=1,2,...,k)$ 是论域 R 上的部分集合, $m_t(i),M_t(i)(t=1,2,...,n)$ 是定义在论域集 $d(i)$ 上的最小和最大值(非负), 即:

$$m_t:d(i)\to[0,+\infty];M_t:d(i)\to[0,+\infty]$$

其中 $m_t(i)\le M_t(i)$

设 $M=\max M_t(i),(1\le i\le k,1\le t\le n)$

令 $t_{A_t}(d_i)=m_t(i)M,1-f_{A_t}(u_i)=M_t(i)M$

其中 $d_i=d(i)$; 则可得出:

$$A=\sum_{i=1}^{k}\frac{[t_{A_t}(d_i),1-f_{A_t}(d_i)]}{u_i}$$

构成了论域 $d(i)(i=1,2,...,k)$ 上的一个区间模糊时序集 A_t. V 是 $A_t(t=1,2,...,n)$ 的集合, V 称为 $d(i)(i=1,2,...,k)$ 的区间模糊时间序列[10].

定义 3. 最佳分段比(w): 最佳分段比是将时序数据进行分段处理的分段值; 分段比小可能处理的效率得不到有效提高, 若太大, 处理的效率会更大;

对于时序数据 $X=\{x_1,x_1,\cdots x_n\}$, 存在分段比将时序能够进行降维生成多个 $N=n/w$ 的子时序数列. 其中 N 为每块子时序数列的大小. 确定最佳分段比对分析数据十分重要, 本文采用 w 对算法效率的影响来确定.

1.2 时序特征的物联网社区

定义 4. 边链接系数: 对于一个无向无权图 $G=(V,E)$, V 和 E 链分别表示图的节点数与边数. 则两节点的边链接系数(Edges Links Coefficient) ELC 可定义为:

$$ELC(v_i,v_j)=\frac{N(v_i)\cap N(v_j)}{N(v_i)\cup N(v_j)},(v_i,v_j)\in V$$

$N(v_i)$ 表示节点 v_i 的关联节点数, 通过共同关联节点与它们所有的关联节点的比值衡量两节点之间的关联程度.

定义5. 模块度: 社区模块度指标 Q 是用于刻画社区特性强弱的参数, 本文采用 Newman 和 Girvan 定义的模块度函数 Q 作为衡量标准, 判断社区划分的合理性和有效性. 定义如下[11]:

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j) \tag{2}$$

其中, k_i 和 k_j 是节点的度值; C_i 是节点 R_i 所属社区; m 是网络总边数. $\delta(C_i,C_j)$ 的计算公式如下:

$$\delta(C_i,C_j)=\begin{cases}1, & C_i=C_j\\ 0, & C_i\ne C_j\end{cases} \tag{3}$$

Q 值在 0 ~ 1 之间, 一般以 $Q=0.3$ 作为网络具有的明显社区结构的下界.

定义6. 物联网社区: 物联网社区就是将具有相同或者相似信息的节点构建成具有关联性的主题社区. 本文定义物联网社区是一个五元组 IC(Internet community) ={ ID, N, Q , R, E }

其中, ID 为社区的编号. N 为社区的主题名称. Q 为社区的模块度, 描述社区特性强弱.

R 为社区成员节点: 它表示其所具有的语义信息(兴趣, 特长责任等(本文以兴趣为例)), 需与具有相同或相似信息的其他节点关联起来. 结合时序数据的特征, 我们定义网络中的节点类似时间顺序进行标号, 即 R_i(i=1,2,3 …) .

E=dis(R_i,R_j)是两个节点之间的边链接系数, 其计算可根据定义 4, 本文中以该节点度通过区间模糊处理后表示, 由于存在所属于不同社区的边缘节点, 其边链接系数不为 0, 即两节点之间有关联性, 根据得出由于共同邻居节点少, 所以比值较小; 我们引入变量 β, 通过 β 来调节边缘节点的划分, β 的值通过实验来调节.

根据时序数据的相关定义, 结合物联网社区的节点特征, 我们将社区节点编号等效为一段时间上的等距时间离散点, 从而将物联网社区划分问题转化在时序数据中进行分类问题. 下面给出基于时序数据的物联网社区的划分算法.

2　基于时序特征的物联网社区划分算法

2.1 时序特征物联网社区划分预处理算法

通过以上相关定义，发现将物联网社区节点数据进行时序化处理后，从时间顺序进行划分，划分的效率会随着社区节点总数增加降低，我们通过最佳分段比 W 对数据进行预处理，下面给出基于分段比思想的预处理算法.

算法: 节点数据预处理算法

Input: $Graph(G) = (V, E)$

//物联网社区的 n 个节点与边的信息(包含各个点的度及下一个邻接点的度信息)

Output: $Graph(G) = (G_1, G_2, G_3 ... G_w)$

//区间模糊、分段处理后的 W 个区域

Begin

For(i=1; i < n; i++)

{　　//分段数为(初始为 3，根据实验调节最佳分段比),对数据节点进行遍历时,

每 n/w 作为一个预处理划分区域;

$V_i = A_i$

//对每段中的节点进行模糊时序处理;

}

End

2.2 物联网社区划分算法

将物联网社区的节点数据进行时序特征化处理后，我们提出的划分算法思想是将离散的节点数据按照时间序列进行遍历，通过边链接系数将离散的节点划分具体的划分算法如下所示.

下面给出该算法的伪代码:

算法: 物联网社区划分算法

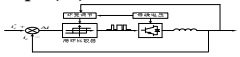
Input:

$Graph(G) = (G_1, G_2, G_3 ... G_w)$

//分段为 W 的物联网节点与边的信息

Output:

$Graph(G)$ 划分处理后的社区 G'

// 

Begin

For (i=1; i < w ; i++)

{　对于 G_i

$ID_1=R_1$; //时序点 1 划为社区 1

m=1; //　社区的数目标记

For(j=2; j<n; j++) //从时序点 2 开始

{

If　($ELC(R_i, R_{i-1}) < \beta$)

将 R_{i-1} 与 R_i 划为同一个社区;

else

for (k=1; k<m; k++)

{

If ($ELC(R_i, ID_j) < \beta$)

将 R_i 划归为社区 ID_j 中;

else

{

以 R_i 为基准点建立新的社区 $D_{(m++)}$;

m++;

}}}

}

2.3 物联网社区划分整合算法

对于上述划分算法，我们得到的是每个分段中划分后的社区，对于整个网络的社区，我们需要对每个分段的社区进行整合调整，下面给出该算法“: .

算法: 社区划分整合算法

Input:

$Graph(G)$ 每个区域进行划分后的社区 G'

// $G' = (G_1', G_2', \cdots, G_w')$

Output:

$G'' = (G_1'', G_2'', \cdots, G_m'')$

//各个区域整合后的物联网综合划分社区

Begin

For (i = 1; i < w ; i++)

For (j = i + 1 ; j < w ; j++)

{ If ($ELC(G_i'(R), G_j'(R)) < \beta$)

{

将 $G_j'(R)$ 划归为 $G_i'(R)$

//对 $G_j'(R)$ 与 $G_i'(R)$ 节点

//进行判断调整

}

End　}

End

2.4 算法效率分析

本文提出的算法主要是首先对节点进行时序化处理，结合模糊集理论，对节点的度量采取区间模糊的边链接系数，通过最佳分段比降低了算法的复杂度. 划分算法通过节点的遍历，判断相邻局部的边链接系

数进行划分，最后进行全局的整合划分. 下面对算法效率进行分析:

表 1 算法的时间复杂性

算法	预处理算法	物联网社区划分算法	物联网社区划分整合算法
时间复杂度	$O(n)$	$O(n)$	$O(n\log n)\sim O(n^2)$

节点数据预处理算法过程中对每个节点进行了一次遍历操作，每个节点遍历过程中主要是对自身信息的初始化，包括节点度的模糊化处理以及相邻节点度的相关处理，并按一定的分段比进行分段，算法的执行的频度为 n，所以其时间复杂性为 $O(n)$；

预处理后的网络，我们采取的是并行计算的思想来提高算法的划分效率. 对每子段进行一次节点遍历，判断相邻局部的边链接系数进行划分，采取局部的划分执行的频度为 n/w，时间复杂度为 $O(n)$；

进行网络整合在每两个子段中进行，算法的时间频度与 w 的取值相关联，分段比 w 的取值至少为 2，则其时间的复杂度范围在 $O(n\log n)\sim O(n^2)$ 之间.

文献[5]与文献[6]通过不断的网络中移除介数最大的边来划分的思想，其时间复杂度文献[5]为 $O(n^3)$，文献[6]为 $O(n^2)$，而本文提出的算法在对网络数据集进行分段处理后，采取并行计算的方式，将时间复杂度降低在 $O(n^2)$ 以内.

3 实验及讨论

为了验证本文提出的物联网社区划分算法方法的性能和效用；本文选取的是社会网络中经典的空手道俱乐部网络数据集. 我们首先利用 Matlab 对算法进行了划分仿真，用 Netdraw 软件对该数据集网络给出了划分后的社区结构；接着通过模块度函数 Q 对网络划分的整体性能进行分析，并与其他算法作出了比较；然后仿真分析了最佳分段比 w 以及 β 对网络社区划分的因素的影响；最后分析了该划分算法与文献[5]及文献[6]算法对比，体现了本文算法的高效性. 仿真实验如下:

3.1 网络划分的结果

通过对 Zachary 通过空手道俱乐部网络社区的划分结果分析，本文算法将该网络划分为 4 个社区，各个社区内节点之间的关联性较好，社区之间的关联较差，宏观上可以看出本文算法能够将网络进行有效划分.

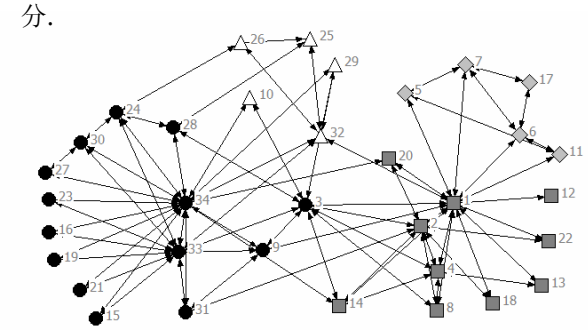


图 1 Zachary 空手道俱乐部网络社区划分结构图

3.2 网络划分的模块度比较

表 2 不同划分算法 Q 的比较

算法	划分社区数	模块度Q	准确率
K-Means	4	0.37	80%
GN算法	3	0.35	72%
Newman快速算法	4	0.38	75%
本文算法	4	0.37	83%

不同算法对网络的划分的结果存在差异，由上表可以得出以下结论: (1)上表所示的算法划分的社区数及模块度 Q 基本相同，即均能将网络进行有效划分; (2)从准确率来看, Newman 快速算法和 GN 算法划分的准确率较低，K—Means 比较高，本文算法比这几种算法准确率均高.

3.3 最佳分段比 w 及 β 的影响

3.3.1 最佳分度比 w 对划分效率的影响

由于所选取的网络数据集节点数不多，我们研究 w 的范围在 2~5 之间，下面给出影响图示:

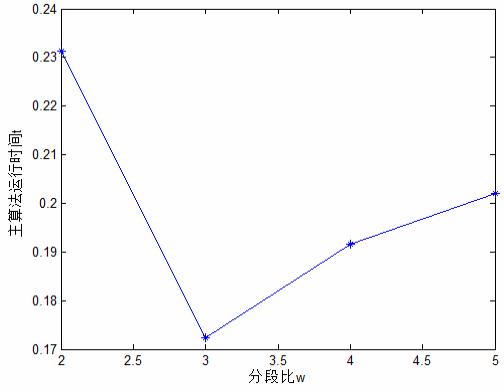


图 2 w 对时间复杂度的影响图

通过对 w 在不同取值下，划分算法的效率进行分析，由上图得出对于该数据集分段比在取时，算法运行时间明显要小，当分段比取小于或大于该值时算法的运行时间达不到最低.

3.3.2 β 对划分结果的影响

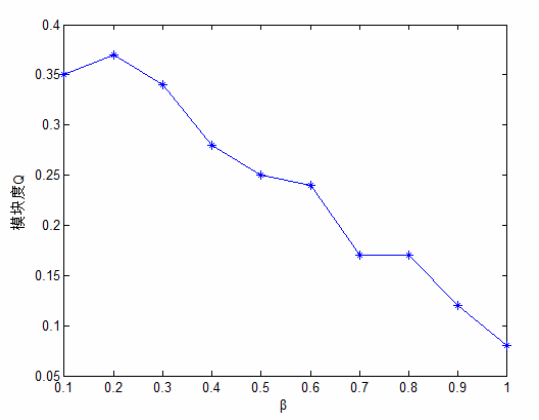


图 3 β 对 Q 的影响图

利用 Matlab 软件对不同的 β 值对社区结构模块度的影响，给出以下结论：(1) β 的取值存在一个最佳值，当其为 0.2, Q 的值达到最大 0.37，即划分的社区结构达到最稳定状态；(2) 从图 4 可以看出 β 的值在 0.2 以后对 Q 的值影响变快，所以 β 对社区的划分影响较大.

3.4 效率比较

表 3 不同划分算法效率比较示意

算法	Kernighan-Lin	GN算法	Newman快速算法	本文算法
	$O(n^2\log n)$	$O(n^3)$	$O(n^2)$	$O(n\log n \sim n^2)$

实验 3.4 将本文算法与其他的几种划分算法进行比较. 本文算法，在处理的数据最佳分段后，能够将时间复杂度降至 $O(n\log n)$ 本文算法在最差情况下时间复杂度为 $O(n^2)$，与其他几种算法相比，在保证准确率的同时，时间复杂度有效的进行了改善.

4 结语

本文针对物联网社区的划分，首先基于时序特征的对物联网社区节点进行模糊化定义，在边链接系数的基础上，结合时间序列的一般特征，提出了一种基于时序特征的物联网社区划分算法；通过仿真验证该算法在能较好的划分社区结构的同时，将划分准确率也有所提高.

参考文献

1 李刚,孙红梅,李智,余海燕,资源受限 Web 服务.计算机学报,2010,33(2):193−206.

2 Huang YH, Li GY. Internet of Things: Semantics, Properties and Category.2011.1.

3 Kernighan BW, Lin S. An efficient heuristic procedure for portioning graph. Bell System Technical Journal, 1970,49: 291−307.

4 Fiedler M. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 1973,23(98):298−305.

5 Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks. Computer and Information Sciences,2005,284−293.

6 Tyler J, Wilkinson D, Huerman B. Email as spectroscopy: Automated discovery of community structure within organizations. International Conferencer on Communities and echnologies, 2003,81−96.

7 Girvan M, Newman NEJ. Community structure in social and biological networks. PNAS, 2001,99(12):7821−7826.

8 Newman NEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6): 66−133.

9 Lin J, Keogh EJ, Lonardi S, Chiu BY. A symbolic representation of time series, with implications for streaming algorithms. Zaki MJ, Aggarwal CC, eds. Proceedings of the 8th SIGMOD Workshop on DMKD 2003. San Diego: ACMPress,2003: 2−11.

10 Gao XD, Cui W, Xu ZY. Similarity matching algorithm for interval-valued time series based on dynamic time warping.1000−5781(2007)06-0664-05.

11 Newman MEJ, Girvan M. Finding community structure in very large netweorks. Physical Review E, 2004,69(2): 26−113.