

一种语义熵的社区划分模型及其应用研究

周明强 朱庆生 刘慧君 张 程

(重庆大学计算机科学学院 重庆 400044)

摘 要 根据复杂网络中整个网络由若干个社区组成和用户通常只对少数主题感兴趣的事实,通过社区语义熵和社区间语义熵,提出了一种基于语义信息的社区结构划分模型,将网络划分为几个语义社区,并将其应用在服务注册中心的具体问题中,同时通过社区负载容量等参数进行了实验分析。实验结果表明,该模型充分考虑到了社区间的语义特性,在应用中效率有显著提高,为语义社区结构中的服务注册中心部署提供了新的途径。

关键词 复杂网络,社区划分,语义熵,服务注册中心

中图法分类号 TP301 **文献标识码** A

Community Division Model by Semantic Entropy and its Application Research

ZHOU Ming-qiang ZHU Qing-sheng LIU Hui-jun ZHANG Cheng

(Department of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract According to the facts that a complex network composed of a number of communities and users are usually interested in only a few topics, this paper proposed a community division model by community semantic entropy and entropy between semantic communities, classed the network into several communities, and applied the model in the services center specific problem. The community load capacity and the other characteristics experimental analyses show that this model's efficiency significantly increases in application, for the semantic characteristics of community are fully accounted, it also provides a new way to the deployment of service registries in semantic community.

Keywords Complex networks, Community division, Semantic entropy, Service registry

1 前言

Watts 和 Barabási 等揭示出很多实际系统构成的网络具有普适的小世界和无标度特征后,复杂网络便成为交叉学科研究的重要内容。其间,对小世界网络和无标度网络的研究多集中在讨论网络模型以及揭示现实网络复杂性特征,对社团结构也多集中在探测方法的研究,但缺乏社团结构研究的系统理论和方法。社团结构分析在社会网络分析中有着重要意义,作为一种观察网络层次结构性质的方法,成为了解网络结构与功能的重要途径^[1,2],近年来,社会社团性质引起了人们的广泛关注。

要准确得到网络中各个节点的准确划分,特别是大规模的复杂网络是一个 NP 难问题^[3]。目前已提出了很多分析复杂网络社团结构的算法,其中著名的算法有 Kernighan-Lin 算法^[4]和基于 Laplace 图特征值的谱平分法(Spectral Bisection Method)^[5,6]算法。另外,北师大小组在已有比较性定义的基础上,提出了适应性自包含比较性定义并设计了相应的最小覆盖算法。段晓东等^[7]提出了一种基于粒子群优化算法的网络社区发现方法。许多试探性算法在一定的条件下可以

得到特定问题的满意解,但是大部分算法还存在其时间复杂度过高等缺陷。

因此,利用语义和社区结构特性来搜寻和发现语义社区(Semantic Community,根据节点兴趣偏好的语义信息划分的内部联系紧密而外部联系稀疏的逻辑结构),然后在其中设置服务注册中心的研究具有重要的实际应用价值,也对复杂网络的理论研究起到了积极的促进作用,在语义社区中进行服务注册中心部署策略研究具有理论价值。

2 语义社区划分模型

现实社会中,在寻求解决问题帮助的时候,通常想到的是领域的专家,以快速得到解决方法,提高解决问题的效率。同样,网络中节点在任何一段时间内,通常只对少数几个主题感兴趣并且尝试连接。由此可见,若将网络中的节点根据自身的兴趣偏好主题组成若干个语义社区,有助于社区内节点快速查询到相应的服务信息,缩短服务响应时间。同时,在语义划分过程中,社区内各节点之间兴趣偏好尽可能相似,而社区间主题差异尽可能地大。

2.1 社区语义熵

系统的状态函数熵是量度系统混乱度的函数,隔离系统

到稿日期:2010-11-15 返修日期:2011-03-25 本文受国家科技部科技支撑计划(2011BAH25B00),中央高校基本科研业务费(CDJZR11180003,CDJRC11180002)资助。

周明强(1977—),男,博士,讲师,主要研究方向为复杂网络、面向服务软件架构,E-mail:zmqmail@cqu.edu.cn;朱庆生(1956—),男,博士,教授,博士生导师,主要研究方向为现代服务业、数字农业和离群数据挖掘;刘慧君(1975—),男,博士,讲师,主要研究方向为 Web 挖掘、社区网络;张程(1977—),男,博士后,主要研究方向为社区网络和移动自组网络。

内的一切可能发生的变化均朝熵增大的方向进行,也就是朝系统混乱度增大的方向进行。在物理学中,熵用来描述原子分布的无序程度:系统的微观状态数越多,热力学概率越大,系统越混乱,熵就越大;当某一系统越有序、越确定时,该系统的熵越小。在信息论中,信息熵由一个信源发出某一消息所包含的信息量来度量,当某一信源发出的消息越确定时,该信源的信息熵越小。

网络中节点的分布类似于原子的分布,当社区的划分越合理,节点在某一社区上的归属越确定时,该社区内的熵值越小。

若 $N=(V,E,F,L)$ 表示一个对等网络,其中 $V=\{v_1, v_2, \dots, v_n\}$ 代表网络模型中顶点的集合, $E=\{e_1, e_2, \dots, e_n\}$ 代表网络模型中边的集合, $F=\{f_1, f_2, \dots, f_n\}$ 表示网络模型中各个顶点的访问频度, f_i 即表示网络中的第 i 个顶点访问服务注册中心节点的频度, L 表示任意两个相邻节点之间的兴趣偏好语义相似度向量的集合。

定义 1 若网络 N 中的节点数为 n , 网络中任意两个节点之间的语义距离向量为 l_i , 其中 $i=1, 2, \dots, n(n-1)$, 设其中存在的 m 个不同的语义距离分别是 $d_1, d_2, \dots, d_m$, 则社区语义熵 S 为:

$$S = -\sum_{j=1}^m P(d_j) \log P(d_j) \quad (1)$$

式中, $P(d_j)$ 为距离为 d_j 的概率, 即 $P(d_j) = \frac{k}{n(n-1)}$, k 为系统中距离为 d_j 的个数。

社区语义熵 S 越小, 表示该社区内各节点间的语义一致性越高, 即社区内节点具有相同的兴趣偏好, 该社区的划分更加合理, 反之, 社区语义熵 S 越大, 表示该社区内各节点间的语义一致性越低, 即社区内节点兴趣偏好差异较大, 该社区的划分不够合理。

2.2 语义社区划分思想

若系统存在 m 个主题, 即网络 N 中存在 m 个不同的语义距离分别是 $d_1, d_2, \dots, d_m$ 。如果不考虑社区的数量, 则可以根据兴趣偏好信息将其划分成 m 个语义社区 $C_1, C_2, \dots, C_m$, 其中 C_m 由节点 $P_i, i=1, 2, \dots, N_m$ 组成。对于一个新的节点 P_x , 根据熵的物理意义, 如果把 P_x 加入到 C_1 后增加的熵小于把 P_x 加入到 C_2 后增加的熵, 则 P_x 应该划分在 C_1 社区中。

一般地, 如果网络已初步划分成了 M 个语义社区: $C_1, C_2, \dots, C_M$, 且满足条件:

$$\begin{aligned} S(C_i + P_i) - S(C_i) &< S(C_m + P_i) - S(C_m) \\ m=1, 2, \dots, M, m \neq i \end{aligned} \quad (2)$$

则 P_x 将被划分到语义社区 C_i 中, 其中 $S(C_m)$ 表示语义社区 C_m 的语义熵。

由于基于语义的社区划分要求社区内各节点之间的兴趣偏好尽可能小, 而社区间主题差异尽可能大, 因此, 基于语义熵的社区划分要求应使得社区内熵尽可能小, 而社区间熵尽可能大, 这样才能满足要求。

若要求划分语义社区的个数小于 m , 则需要语义社区合并。因此, 参考文献[8]定义两个语义社区之间的社区间熵, 用于确定语义社区的数目:

$$S(C_i, C_j) = -\log V(C_i, C_j) \quad (3)$$

式中, $V(C_i, C_j) = \frac{1}{(2\pi)^n} \sum_{i=1}^N \sum_{j=1}^N M(d_i, d_j) \exp$

$$\left(-\frac{(d_i - d_j)^T (d_i - d_j)}{2h^2} \right), M(d_i, d_j) = \begin{cases} 1, & d_i \in C_i, d_j \in C_j \text{ 或} \\ & d_i \in C_j, d_j \in C_i \\ 0, & \text{其它} \end{cases}$$

h 是窗宽或平滑系数。

如果社区 C_i 和 C_j 分得比较开, 则 $V(C_i, C_j)$ 比较小, 从而 $S(C_i, C_j)$ 比较大; 反之, $V(C_i, C_j)$ 比较大, 而 $S(C_i, C_j)$ 比较小。

语义社区合并操作具体步骤如下: 计算任意两个社区间的社区间熵, 合并社区间熵最小的两个社区, 重复这个步骤, 直到得到需要的社区数量。

要准确得到网络中各个节点的准确划分, 特别是大规模的复杂网络是一个 NP 难问题, 如果在某个时刻计算的两个类最小的社区间熵发生突变, 则表明网络由无序到有序以及由有序到无序的突变, 此时各社区分得最开, 社区的总数可以作为最佳语义社区数目的一个参考。当然也有很多其它的合并的算法, 本文不再做详细的讨论。

3 语义社区分析

在语义社区确定后, 可以利用超级节点选择算法^[9]为每个确定的语义社区选择其中的一个节点作为社区内服务注册中心, 承担社区内部普通节点的服务注册和查询。若节点无法在所属社区内部服务注册中心查询到所需要的服务信息, 则由社区内服务注册中心将服务请求路由到其它社区中的服务注册中心节点服务器上注册和查询等服务。

在社区划分的过程中, 社区内部的连接数决定网络运行的效果。社区内连接数过多, 使得社区内服务注册中心的负载增加, 从而出现单点故障等问题; 社区内连接数过少, 社区内服务注册中心的服务信息偏少, 服务请求需要频繁地路由到其它社区中的服务注册中心上, 造成系统整体服务响应时间的下降。因此在社区划分过程中的社区间负载均衡反映了社区内部连接的变化情况。

3.1 社区性能参数

定义 2 若 C_p 表示网络中的某个语义社区, v_i 表示社区内部的节点, f_i 表示节点 v_i 的访问频度, 则社区内部负载容量 LD_{C_p} 表示为:

$$LD_{C_p} = \sum_{v_i \in C_p} f_i \quad (4)$$

当社区内部的服务注册中心部署后, 社区中的连接主要体现在社区内部节点与服务注册中心之间的连接上, 因而服务注册中心的负载表现为在单位时间内接收到的服务连接请求数量之和, 即 $LD_{C_i} = \sum_{v_i \in C_p} f_i$, 所以, 社区内部负载容量可以视为在社区划分后社区内部的负载。

定义 3 社区负载均衡指数 D 表示的是各个社区内部负载容量之间差距的大小, 即:

$$D = \frac{1}{m} \sum_{C_p \in V} (LD_{C_p} - \frac{1}{m} \sum_{C_q \in V} LD_{C_q})^2 \quad (5)$$

式中, m 表示整个网络中的社区数, C_p 与 C_q 表示网络中的社区, LD_{C_p} 与 LD_{C_q} 表示社区 C_p 与 C_q 对应的社区内部负载容量。 D 值越小, 则说明各个社区之间的负载容量差距越小, 即社区间的负载越均衡。

服务注册中心是社区的中心节点, 本文引入社区内部节点最短网络时延指数 $T_{\infty}(v_i)$ 来衡量社区内部任意节点到节点 v_i 的最短网络时延。

定义4 节点 v_s 的最短网络时延指数 $T_{as}(v_s)$:

$$T_{as}(v_s) = \sum_{v_j \in V_p} t(v_j, v_s) \times f_j \quad (6)$$

$t(v_j, v_s)$ 表示社区 C_p 中节点 v_j 到 v_s 的最短网络时延。 $T_{as}(v_s)$ 越小,说明社区中的节点到 v_s 的访问时间越短。在后续部署服务注册中心时,将节点的最小 $T_{as}(v_s)$ 值作为部署是考虑的重要因素之一。

3.2 社区性能计算

在社区中节点的性能参数等信息通过消息传递进行传递,节点间根据这些消息计算网络时延等。考虑到划分后的语义社区一般规模不大,而这些参数信息是必须传递的信息,这里消息传播采用简单的 Flooding 方式。

通知消息 $Meg_{notify} = \{N_{id}, M_{id}, TTL, t_{notify}, f\}$, 其中 N_{id} 表示节点的 ID 号; M_{id} 表示消息 ID, 它通知消息的唯一标识; TTL 表示消息生存时间; t_{notify} 表示消息的发送时间。

社区中的每一个节点设置节点 v_s 所拥有的最短网络时延结构数组 TA_s , 记录社区内部其它节点到达节点 v_s 的一个最短网络时延。节点 v_s 通过向社区中的其它节点广播消息 Meg_{notify} , 使得各个节点获取到到达其它节点的网络时延。

在社区中节点的性能参数等信息通过消息传递进行传递的具体算法描述如下。

算法1 BroadCast_TimeMeg

输入: 节点 v_i 编号和网络 N ;

输出: 网络状态 N ;

算法的步骤:

- 1) for $\forall i \in [1, n], n \in N$
 - {
 - 创建并广播 v_i 通知消息 Meg_{notify} ;
 - 2) 节点 v_s 收到 Meg_{notify} :
 - {
 - 3) $T_{as}(v_s) = \sum_{v_j \in V_p} t(v_j, v_s) \times f_j$; // 计算 $t(v_j, v_s) \times f_j$;
 - 4) Update TA_s ; // 更新 v_s 最短网络时延数组;
 - }
 - 5) if $TTL > 0$ then
 - {
 - 6) $TTL = TTL - 1$;
 - 7) Goto(2);
 - }
 - // TTL 减 1 后消息继续发送跳转步骤 2);
 - }

4 语义社区划分算法

每个节点设置一个社区内部结构数组 AL_i 存储网络中各个语义社区的连接状态等信息, 其中 n 表示整个网络中的节点个数。语义社区结构数组中的每一条记录表示网络中一个社区内部连接信息, 其包括节点 ID 号 N_{id} , 社区 ID 号 C_{id} , 社区内部负载容量 S 和与该社区相连接的社区的 ID 号集合 $LinkCIDs$ 。若该节点为服务注册中心节点, 则记录中的 N_{id} 与 C_{id} 相同。在下一步研究中, 可以进一步优化结构, 并结合语义路由的相关内容, 将其作为语义路由表的重要内容。

基于社区划分的服务注册中心节点部署 CSPS (Community-based Service Peer Selection, CSPS) 具体算法如下。

算法2 Semantic_Community

输入: 初时网络 N ;

输出: 社区结构和社区的节点设置。

算法的步骤:

- {
- 1) 初始化网络 N , 创建并广播通知消息 $COMMUNITY_MSG$;
- 2) for $\forall i \in [1, n], n \in N$
- 3) {
- 4) Create AL_i ; // 创建所属社区结构数组;
- 5) $m \leftarrow \text{Count}(d_i)$ // 统计语义值个数;
- }
- 6) if $H(C_p + P_i) - H(C_i) < H(C_p + P_i) - H(C_k)$ then // 若社区划分;
- {
- 7) $C_p \leftarrow C_p + C_q$; // 修改社区 V_p 的相关信息;
- 8) $LinkCIDs_{V_p} \leftarrow LinkCIDs_{V_p} + LinkCIDs_{V_q}$; // 修改与 V_p 的 $LinkCIDs$
- 9) Delete_Community(V_q); // 删除社区 V_q ;
- }
- 10) for $i = 1$ to m ;
- 11) Select_Super_Peer;
- // 利用超级节点选择算法^[9]设置每个社区中的服务注册中心节点;
- }

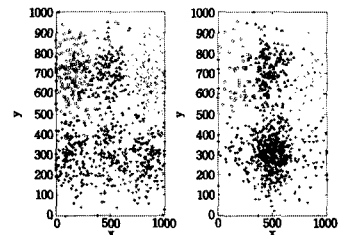
算法实现了网络语义社区的动态划分和每个社区服务注册中心超级节点的动态选择。网络中的每个节点获得通知消息 Meg_{notify} 后, 节点可以独立计算并完成合并的操作, 由于各个节点都设置相同的社区负载向量数组, 因此每个节点最后都会得到相同的社区划分结果。

5 实验与结果分析

实验采用 MATLAB 仿真, 以节点数 $n = 1000$, 连接概率 $p = 0.5$ 生成的随机网络作为实验数据。实验中连接频度值采用了服从 $N(50, 10)$ 的正态分布和服从 $U(0, 100)$ 的均匀分布。为了简化实验, 本文在构造数据集的语义信息时, 将语义主题信息设置为 6 类。

5.1 社区结构分析

社区结构划分对比分析: 实验主要目的是为了对比分析语义社区划分 CSPS 算法与 SCP2P 算法对网络社区结构划分的影响情况。其具体对比情况如图 1 所示。



(a) CSPS 算法 (b) SCP2P 算法

图1 社区结构划分的影响情况

图 1(a) 显示出了 CSPS 算法对初始网络的划分情况, 即在社区数量为 6 时的划分情况。图 1(b) 显示出了 SCP2P 算法在指定社区数量为 6 时的划分情况。

从图 1 中可以看到, SCP2P 算法以绿色顶点和紫色顶点为代表的两个社区中的顶点数量明显多于其它社区中顶点的数量。而在图 1(a) 中, 各个社区之间的节点数量差距不大。因为实验初始情况下, 各个节点的 w 服从正态分布, 节点 w 值差别很小, 社区之间的差异度可用社区之间的节点数目近

似来比较。所以,可以得出 DSPS 算法在社区划分的负载均衡指数上优于 SCP2P 算法。

5.2 社区间负载均衡指标

社区间负载均衡指标分析:实验目的以社区划分过程中社区之间的负载均衡程度 D 作为衡量指标研究网络节点初始状态对算法的影响。图 2 显示了初始网络中节点的连接频度服从不同的分布,对社区之间 D 指标的影响程度。

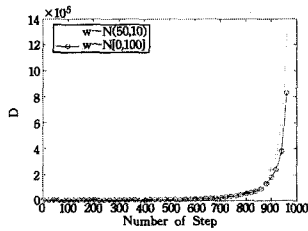


图 2 网络初始状态对算法的影响

从图 2 中可以看到,两种节点频度的选择方法产生的两条曲线比较接近,说明 CSPS 模型对初始网络中节点频度服从的两种分布都有很好的适应性。

图 3 与图 4 分别表示 CSPS 模型、SCP2P 模型及 PDC 模型在负载均衡指数 D 上的变化情况。

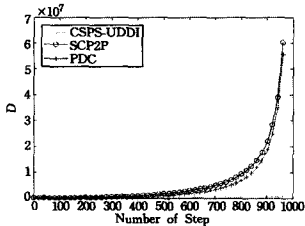


图 3 w 服从正态分布时 D 的变化情况

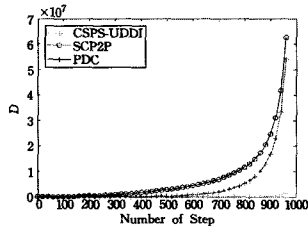


图 4 w 服从均匀分布时 D 的变化情况

从图 3 和图 4 可以看到, CSPS 模型、SCP2P 模型、PDC 模型在步骤增加的过程中,负载均衡指数 D 都会增加,且社区数目越少,负载均衡指数 D 越大。对于正态分布和均匀分布连接频度的初始网络, CSPS 模型在性能上明显优于 SCP2P 模型和 PDC 模型,其主要原因是 CSPS 模型在合并社区的时候,以兴趣偏好差异最小为合并依据。

5.3 平均服务查询时间

平均服务查询时间分析:社区中的服务注册中心负责整个社区网络中的服务注册和查询等服务,其最重要的指标是网络中的各个节点连接服务注册中心的注册和查询服务的响应时间。

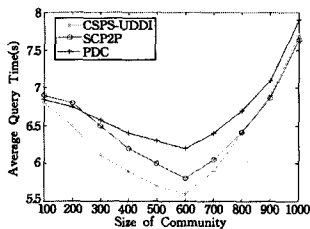


图 5 不同算法平均查询时间的对比情况

本实验的目的是比较 CSPS 模型、SCP2P 模型和 PDC 模型在社区节点数目增长的情况下,平均查询时间的变化情况。

从图 5 可以看到,在社区数量接近 600 时,3 种模型都取得了最小的平均查询时间,但是 CSPS 模型具有更低的平均查询时间。CSPS 模型部署社区内部的语义熵最小,即保证节点查询集中在语义内部完成,从而保证在每个社区内部,该服务注册中心节点的部署都是最优的。

但是,随着社区数目的增加,社区内部节点往往不能在内部服务注册中心查询到所需要的信息,需要路由到其它社区中,增加了社区间的路由时间,而该实际过程一般比内部访问所花要多得多,从而增加了节点的平均查询时间。当社区内部节点数比较多时,选择社区内部平均网络时延往往不能有效地减少节点的平均查询时间,所以,当社区内部节点数比较多时, CSPS 模型与 SCP2P 模型比较接近。

结束语 本文的主要思想是将网络基于主题领域划分为若干个语义社区,通过引入社区语义熵和社区语义间熵,提出了一种基于社区语义熵的社区划分方法,并将其应用到服务注册中心的具体应用中;通过社区负载容量等参数对以语义社区划分为主要依据的服务注册中心节点部署模型 CSPS 模型进行了实验。实验结果表明,采用语义社区的思想,在每个语义社区内设立服务注册中心节点统一管理每个社区服务信息,社区划分后,由超级节点管理每个语义社区,并承担服务注册中心任务,节点根据语义信息自主地建立并维护与其它节点的拓扑连接,完成网络自组织过程,以达到节点资源的有序组织,有利于提高社区内的访问效率,从而减少社区间的路由访问,减少网络带宽占有量。因此在网络中自动搜寻或发现社区,然后在其中设置服务节点的研究具有重要的实际应用价值。

参考文献

- [1] 李晓佳. 复杂网络中的社团结构[J]. 复杂系统与复杂性科学, 2008, 5(3)
- [2] 汪小帆. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报, 2009, 38(5)
- [3] 解邹. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学, 2005, 2(3)
- [4] Kernighan B, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291-307
- [5] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(2): 298-305
- [6] Pothen A, Simon H, Liou K. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11: 430
- [7] 段晓东. 基于粒子群算法的 Web 社区发现[J]. 计算机科学, 2008, 35(3)
- [8] Gokcay E, Principe J C. Information theoretic clustering[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(2): 158-171
- [9] Zhou M, Zhu Q, Li S. Research on forecast and estimate of super-node service capacity[J]. Journal of Information and Computational Science, 2009, 6(2): 941-950