论 文

语义对等覆盖网中社区结构的发现和评价

陈汉华, 金海*

华中科技大学计算机学院服务计算技术与系统教育部重点实验室、集群与网格计算湖北省重点实验室, 武汉 430074

* 通信作者. E-mail: hjin@hust.edu.cn

收稿日期: 2010-05-13; 接受日期: 2011-04-04

国家自然科学基金委员会 (NSFC) 与香港研究资助局 (RGC) 联合资助基金 (批准号: 60731160630)、高等学校博士学科点专项科研基金 (批准号: 20110142120080) 和国家自然科学基金 (批准号: 61003006) 资助项目

摘要 最近, 通过建立语义覆盖网络来提高大规模分布式网络环境中信息检索服务的性能成为对等计算领域的研究热点. 目前, 研究者在语义覆盖协议和搜索算法方面已经做了大量研究, 证明了语义覆盖在基于对等网络模型的内容定位应用方面极为有效. 然而, 分析和评价语义覆盖网络特征的研究工作确非常有限. 文中通过建立数学模型和设计启发式回溯-贪婪混合算法, 确认了语义覆盖网络的一种主要内在特性——社区结构特性. 利用评价模型比较了 SemreX 语义覆盖网络和 Gnutella 网络的性能, 实验结果显示 SemreX 覆盖网具有显著的社区结构特征, 而 Gnutella 网络却没有这样的特征. 另外, 通过分别在两种覆盖网中仿真洪泛协议发现具有显著社区结构特征的覆盖网在内容定位方面效率更高.

关键词 对等网络 社区结构 语义覆盖 语义相似度 搜索效率

1 引言

自从对等 (peer-to-peer, 简称 P2P) 文件共享应用, 例如 Napster 和 Gnutella 兴起以来, 数百万的网络用户开始使用 P2P 工具来获取想要的文件. P2P 网络作为网络信息获取工具显示出巨大的潜能^[1].

目前 P2P 文件共享领域的工作主要研究基于文件名称的数据共享. 现有文件共享系统按照结构特征可以分为 3 类^[2]. 第 1 类是基于集中索引的 P2P 结构, 所有搜索请求经过集中索引服务器查询处理后才能定位并访问资源. 显然, 随着 P2P 网络规模的扩大, 集中索引服务器必然成为系统性能的性能瓶颈和单一失效点. 第 2 类 P2P 结构基于分布式哈希表 DHT^[3,4] 技术. DHT 具有良好的可靠性和可扩展性, 但由于其哈希映射基于 (key value) 对的精确匹配特性, 难以高效支持复杂的内容检索^[5]. 虽然已经有少量基于 DHT 的系统提供部分匹配查找功能的工作^[6~8], 但 DHT 固有的特征使研究的方法非常复杂, 且远不如现流行的 Web 搜索引擎有效. 第 3 类通常被称为无结构的覆盖 (unstructured overlays). 这类系统对 P2P 网络的拓扑结构不进行限制, 主要采用洪泛或随机游走的搜索策略^[9]. 一般来说, 洪泛搜索策略会使查询消息在 P2P 网络上以指数的方式增长进而导致严重的网络拥塞, 这使得 P2P 网络的可扩展性受到限制; 而随机游走虽减少了网络流量, 但一般牺牲了系统查全率, 且查询

引用格式: 陈汉华, 金海. 语义对等覆盖网中社区结构的发现和评价. 中国科学: 信息科学, 2012, 42: 537–549

处理具有较大的延迟, 系统可靠性难以得到保证. 可见, 传统的 P2P 覆盖网络结构无法有效支持大规模 P2P 内容检索.

最近, 研究者们已经将注意力转向创建语义覆盖网络 (SON) 来提高大规模 P2P 网络平台下信息检索的效率. 著名的例子有兴趣覆盖^[10], Edutella^[11,12], Bibster^[13] 和 SSW^[14]. 在文献 [10] 中, Sripanidkulchai 等发现了在 Gnutella 及其他流行 P2P 网络中都拥有的用户行为特性, 被称为兴趣局部性原理, 这个原理指出如果某结点有一个用户感兴趣的文件, 那么此结点很可能有其他类似的文件. 基于此原理, Sripanidkulchai 等在 Gnutella 基础上提出并实现了一种自组织协议, 将那些拥有相似兴趣内容的结点连接起来, 使得结点可以通过这些兴趣捷径来定位感兴趣的内容, 大量减少网络中的洪泛消息. Edutella^[11,12] 使用 RDF 元数据模型来标准化数据的组织方式并且基于其使用的元数据模式、属性、本体来实现路由和结点聚集策略. 在 Bibster^[13] 中, 所有结点在网络中广播自己的专有知识即它的主题集合, 其他收到这些广播消息的结点与自己的邻居结点创建一个语义链接, 所有语义链接形成一个语义覆盖网, 并以此作为查询消息智能路由的基础. SSW^[14] 覆盖网络根据数据的语义相似度动态地聚集网络结点, 使语义相似度高的结点从逻辑上更紧密地联系在一个组群内, 并将结点集群由高维的语义空间映射成一维小世界空间, 这种做法能有效地使系统在搜索路径长度和维护成本之间达到平衡.

可见, 研究者们在设计语义覆盖协议和搜索算法方面已经做了大量的研究工作, 并且证明语义覆盖策略对于内容搜索非常有效. 然而, 在分析和量化语义覆盖网络特征方面, 目前的工作却非常有限. 众所周知, 大规模网络系统往往非常难以管理, 而对网络拓扑属性的分析往往有利于权衡系统设计的利弊^[15]. 本文认为语义覆盖网络仍须更深入的实证和理论建模工作以更好地理解这种覆盖网络模型, 进而据此设计更强健的搜索协议来提高 P2P 信息共享系统检索性能, 同时减少网络中不必要的消息提高网络带宽资源的利用率.

本文发现了语义覆盖网络的一个显著特征——社区结构^[16], 即语义 P2P 中内容相似性高的结点被聚集在一起, 网络的结点自组织成一些组群, 一个组内的结点连接紧密, 组之间的连接比较松散 (图 1). 基于 SemreX 系统——一个由华中科技大学集群与网格计算实验室开发的用于文献共享的语义 P2P 平台, 来阐述社区结构的性质, 建立了评估语义覆盖网社区结构特征的数学模型, 并设计了启发式回溯-贪婪混合算法用于模型求解. 通过比较 SemreX 语义覆盖网络和 Gnutella 网络的性能, 结果显示 SemreX 覆盖网具有显著的社区结构特征, 而 Gnutella 网络却没有. 另外通过分别在这两种覆盖网中模拟洪泛协议发现具有显著社区结构特征的覆盖网在内容搜索方面更有效率.

本文的主要贡献概括如下: 1) 提出一种新颖的数学模型用来评价语义覆盖网络的社区结构. 据作者所知, 本文是第一个研究语义覆盖网络的社区结构的工作. 2) 设计了一种启发式回溯贪心算法用于求解社区结构模型. 3) 证明了 SemreX 系统中的社区结构特性.

论文结构如下: 第 2 节介绍 SemreX 语义覆盖, 第 3 节提出社区结构评价模型, 第 4 节介绍模型启发式求解算法, 第 5 节实验比较 SemreX 语义覆盖和 Gnutella 网络的社区结构特征, 第 6 节介绍相关工作, 第 7 节总结全文介绍未来工作.

2 SemreX 对等覆盖网

2.1 SemreX 体系结构概述

SemreX 是华中科技大学集群与网格计算实验室研究开发的一种 P2P 文献检索系统^[17]. SemreX

图 1 社区结构

Figure 1 Community structure

图 2 SemreX 系统模型

Figure 2 Four-layer model of SemreX

系统旨在让计算机领域的研究者共享他们文件系统中的科技文献, 搜索和下载他们想要的科技论文. 图 2 说明了 SemreX 系统的体系结构, 包括逻辑概念层, 文档对象空间层, 网络结点通信层, 语义覆盖层这四个层次. 在本系统相关的早期文献 [18,19] 中已经给出了 SemreX 系统体系结构的详细描述, 这里只做简单的回顾.

SemreX 的逻辑概念层的目标在于在分布式 P2P 网络环境下提供共享的概念模型, 用于文献分类及元数据描述. 例如, 计算机科学的文献检索应用中, ACM topic 本体¹⁾被广泛使用, 已成为计算机科学类文献分类的标准本体. 它涵盖了计算机科学主题的 1475 个研究领域. 由于 topic 间具有父子关系, 概念之间以 IS-A 的层次关系被关联起来. SemreX 使用 ACM topic 作为文献分类标准. 作者、标题、出版商等元数据信息是根据通用的元数据模型抽取出来. 文档对象层负责从文档中抽取元数据, 并将文献文档分类到不同的 topic [19]. 在 SemreX 中, 根据属于不同 topic 的文档间的语义相似度, 结点自组织成一个语义覆盖, 查询信息会被路由到与其语义相似的结点以减少冗余的信息传输. 网络结点通信层在整个系统中扮演“传输层”的角色, 它封装了所有底层数据交换的细节 [20].

2.2 语义覆盖

在介绍社区结构评价模型之前, 先给出结点间语义相似性的形式化定义.

定义 1 (结点的语义相似性) 在 SemreX 系统中, 结点间的相似性被定义为相应带权 topic 集合之间的语义相似性, 其中 topic 为 ACM topic 概念树结构中的概念结点.

$$\text{Sim}(P_1, P_2) = \text{Sim}(\{\langle T_i, \lambda_i \rangle, i = 1, 2, \dots, m\}, \{\langle T_j, \lambda_j \rangle, j = 1, 2, \dots, n\}), \quad (1)$$

其中, P_1 和 P_2 表示任意的结点对, T_i 和 T_j 分别表示 P_1 和 P_2 结点中存储的文献被归入的 ACM topic 类别; λ_i 是被归入 T_i 的文档数量在 Peer 文档集合中的统计权重 λ_i 通过下式计算:

$$\lambda_i = \frac{N_i}{\sum_{j=1}^n N_j}. \quad (2)$$

1) ACM topics. <http://www.acm.org/class/1998>

概念间的语义相似度的研究是自然语言处理领域的一个研究问题, 近 10 年来研究者们已经开发了很多语义相似性的度量方法. 其中 Yuhua 的度量方法^[21] 具有较好的效果:

$$\text{Sim}(T_i, T_j) = f_1(l)f_2(h) = \begin{cases} e^{-\alpha} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, & \text{if } T_i \neq T_j, \\ 1, & \text{if } T_i = T_j. \end{cases} \quad (3)$$

这里, l 表示主题之间最短路径的长度, h 表示从 T_i 和 T_j 的子主题结点 (subsumer) 到主题概念的共同祖先结点的最大深度. 参数 $\alpha > 0, \beta > 0$ 用于衡量最短路径长度 l 和深度 h 对语义相似性的贡献. 根据实验的经验, 公式中 α 和 β 的最佳取值为 0.2 和 0.6.

基于以上公式, 给出以下衡量两个结点之间的语义相似性的公式:

$$\text{Sim}(P_1, P_2) = \sum_{j=1}^{|P_2|} \sum_{i=1}^{|P_1|} [\text{Sim}(T_i, T_j) \times (\lambda_i, \lambda_j)]. \quad (4)$$

公式中, $|P_1|$ 和 $|P_2|$ 分别表示两个结点中 topic 的数量, 如公式所表示, 结点之间的相似性是由属于两结点中的任意两 topic 的相似性的所有排列之和求得.

SemreX 语义覆盖的基本想法就是将有相似主题 (topic) 概念的结点连接起来. SemreX 在 Gnutella 协议的基础上设计了一个语义链生成协议, 当结点之间的相似度达到一个预先设置的阈值时, 它们之间将形成一个语义链. 那么在 SemreX 中搜索时, 语义链接覆盖将会引导到最可能有需要结果的结点的路径^[18].

2.3 SemreX 的社区结构

图 3 是一个包含 200 个结点的 Gnutella-like 网络拓扑图. 语义覆盖网则是通过在 Gnutella-like 网络上仿真 SemreX 协议而得到.

研究表明 KaZaa 系统结点的文件数遵循 Zipf 分布^[15], 在 Gnutella-like 拓扑上按照 Zipf 分布模拟文件分布, 且每个文档属于 ACM topic 树中的一个 topic 结点. 然后, 在 Gnutella-like 拓扑的基础上做了语义链接协议 (在文献 [18] 中提到) 仿真, 最终得到 SemreX 覆盖网络. 图 4 是一个语义覆盖网络的例子, 其中结点通过语义覆盖协议重新自组织成一个个组, 在每个组内, 结点是紧密连接的, 而组之间的结点之间是松散连接的, 图 4 同时也显示了邻居结点之间的相似度值, 不难发现 SemreX 覆盖网络中同一个组中的结点间的相似性比不同组中结点高得多, 也就是说, 一个组中结点由于有较高相似性值而拥有更强的语义连接, 组之间的语义连接就弱得多.

3 社区结构的评价模型

众所周知, 网络建模方法已成功应用于描述 Internet、WWW、文献引用等网络, 对优化网络系统的性能提供了指导. 本节用理论化建模方法来评价语义覆盖中社区结构特征, 希望最终能系统开发者更充分利用语义覆盖网络的拓扑特性设计高效的算法. 这里先给出语义覆盖的形式化定义如下:

定义 2 (语义覆盖) 语义覆盖网络被定义成一个网络图 $G = (E, V)$, 其中 E 表示结点之间语义连接的集合, V 表示结点的集合.

用如下的布尔邻接矩阵 B 简单表示网络中任意两结点的连接关系.

图 3 社区结构

Figure 3 Gnutella-like overlay network with 200 nodes

$$b_{ij} = \begin{cases} 1, & \text{if } e_{ij} \in E, \\ 0, & \text{if } e_{ij} \notin E. \end{cases} \quad (5)$$

一个由 n 个结点组成的语义覆盖网络能用一个 $n \times n$ 的邻接矩阵 $\mathbf{A}$ 表示, 其中元素 a_{ij} 不是简单的布尔值 0 或 1, 而是两结点语义连接的权值, 即结点间的语义相似性值.

$$a_{ij} = \begin{cases} \text{Sim}(P_i, P_j), & \text{if } e_{ij} \in E, \\ 0, & \text{if } e_{ij} \notin E. \end{cases} \quad (6)$$

图 4 中不难发现语义覆盖网中结点自组织成社区, 社区中的连接紧密且强度大, 而社区间的连接松散且强度很弱. 下面将说明如何评价这种结构特征的强弱.

假设覆盖网络已经被划分成若干社区, 其中 C_k 表示第 k 个社区. 对称矩阵 $\mathbf{C}$ 的每个元素 c_{mn} 表示社区 C_m 和 C_n 这间的语义连接强度, 即社区 C_m 和社区 C_n 之间所有结点间的连结形成的边的权值总和所占全网边的权值总和比例的一半, 全部比例之和为 $c_{mn} + c_{nm}$.

$$c_{mn} = \frac{\sum_{v_i \in C_m} \sum_{v_j \in C_n} a_{ij}}{\sum_i \sum_j a_{ij}}. \quad (7)$$

唯一例外的是对角线元素 c_{mm} , 它的值等于社区 C_m 中所有边的权值之和占全网所有边的权值之和的比例.

同样构造对称矩阵 $\mathbf{C}'$, 其中的每个元素 c'_{mn} 计量社区 C_m 和 C_n 之间的布尔连结强度, 即社区 C_m 到社区 C_n 的边数之和占全网边数之和的一半, 那么所有连结的边数之和为 $c'_{mn} + c'_{nm}$. 唯一的例外是对角元素 c'_{mm} , 它的值应为社区 C_m 中边数之和占全网所有边数总和的比例.

图 4 SemreX 系统模型

Figure 4 Community structure of semantic overlay network

$$c'_{mn} = \frac{\sum_{v_i \in C_m} \sum_{v_j \in C_n} b_{ij}}{\sum_i \sum_j a_{ij}}. \quad (8)$$

假设所有的连接都断开成两个端点, 令 δ_k 表示与社区 C_k 中的结点相连的端点之和占全网端点总和的比例, 容易得到 $\delta_k = \sum_l c'_{kl}$. 如果所有连接端点按照随机方式连接, 那么这些边落在社区 C_k 内的概率为 δ_k^2 .

社区结构评价模型如下:

$$Q = \sum_k \left(c_{kk} - \delta_k^2 \times \frac{\varpi'_k}{\varpi_{\text{total}}} \right), \quad (9)$$

其中, ϖ_{total} 表示语义覆盖中所有边的平均权值, 即平均相似强度

$$\varpi_{\text{total}} = \frac{\sum_j \omega_j}{|E|}, \quad (10)$$

ϖ'_k 则衡量社区 C_k 内边的平均权值, 包括两个端结点都属于社区 C_k 的边和只有一端属于社区 C_k 的情况.

$$\varpi'_k = \frac{\sum_{e_j \in E_k} \omega_j + \frac{1}{2} \sum_{e_t \in E'_k} \omega_t}{|E_k| + \frac{1}{2} |E'_k|}. \quad (11)$$

其中, E_k 表示社区 C_k 内的边的集合, E'_k 表示只有一端属于社区 C_k 的边的集合. 那么, $\varpi'_k / \varpi_{\text{total}}$ 将社区 C_k 中的边的权重相对与整个覆盖网络的边的平均权重的相对值.

因此, $(c_{kk} - \delta_k^2 \times \frac{\varpi'_k}{\varpi_{\text{total}}})$ 是社区 C_k 中所有边的权值之和占全网权重的比例减去随机连结所得到的比例. 这里用 Q 衡量社区结构的强度, 它是给定社区划分中的所有社区的这种偏差的总和. 不难看出, 如果一种划分与随机划分毫无差异, 那么模型的估算值将接近最小值 $Q = 0$. 大于 0 的值表明不是随机地划分, 此模型的目标是使 Q 值最大化. (9) 式中提出的模型与 Newman 等在文献 [22] 中定义的模型显然不同, 因为 Newman 的模型只考虑不带权的网络.

显然, 模型中 Q 值越高, 对应的社区划分越好. 要评价社区结构的强度, 只需优化社区划分求 Q 的最大值. 问题是, 如果穷尽所有可能的划分, 对一个 n 结点的网络, 其复杂度至少是指数级别的, 穷举计算将非常费时, 这对于大规模系统来说是不切实际的.

4 社区结构划分算法

一般来说, 各种流行优化方法, 如模拟退火算法、贪心算法等都可以解决这种问题. 然而, 在社区划分问题上这些算法不够强大. 例如, 考虑贪心算法, 它的初始状态是每个单一的结点为一个社区, 然后重复将其中的一对社区合并起来, 每一步合并贪心算法都从很多种选择中执行能最大化增加 Q 值的合并. 合并按此执行, 直到 Q 值不再增大为止, 即得到最大的 Q 值和对应最优的划分方案. 这种算法虽然能够较快地找到一个较高的 Q 值, 但是很明显贪心算法的社区合并过程是不可逆的, 往往局部的最优并不意味着全局的最优, 甚至一个不好的开始就决定了一个不好的结果.

针对贪心算法的问题, 本文设计了一种回溯 – 贪心混合算法. 算法的初始状态同样是每个结点自成一个社区. 与标准贪心算法的区别在于每次合并时选择合并前, 先将待选合并方案按贪心算法执行至 Q 值不再增大, 然后从所有待选合并方案中选择能获得最大贪心结果的组合. 可见, 相比标准贪心算法, 这里不是简单执行单步最优, 而是每一步都往后看了若干步, 选择了多步最优.

算法依此执行, 即使 Q 没有增加算法也不会停止, 直到所有的结点被加入到一个社区才会形成最终的划分. 算法 1 的伪代码详细描述了寻找最优社区划分的基于贪心搜索的回溯算法, 算法 2 则给出了每步选择所执行的贪婪搜索子算法.

Algorithm 1 Backtrack based greedy search algorithm

```

/*  $\Gamma$ : the current state of the partition */
backtrace_search( $\Gamma$ ) {
   $Q \leftarrow \text{calculate } Q_{\Gamma}$ ;  $\text{Best\_}Q \leftarrow 0$ ;
   $\text{Best\_}\Gamma \leftarrow \Gamma$ ;  $n \leftarrow \text{number of communities of } \Gamma$ ;
  repeat
     $\Gamma' \leftarrow \Gamma$ ;  $Q_{\max} \leftarrow 0$ ;
    for (each pair  $(C_i, C_j)$  within  $\Gamma$ ) do
       $\Gamma \leftarrow \text{join\_community}(C_i, C_j)$ ;
       $Q_{ij} \leftarrow \text{greedy\_search}(\Gamma)$ ;
      if  $Q_{ij} > Q_{\max}$  then
         $Q_{\max} \leftarrow Q_{ij}$ ;  $i^* \leftarrow i$ ;  $j^* \leftarrow j$ ;
      end if
       $\Gamma \leftarrow \Gamma'$ ; /*backtrace to  $\Gamma^*$ */
    end for
     $\Gamma \leftarrow \text{join\_community}(C_{i^*}, C_{j^*})$ ;
     $Q \leftarrow \text{calculate } Q_{\Gamma}$ ;
    if  $Q > \text{Best\_}Q$  then
       $\text{Best\_}Q \leftarrow Q$ ;  $\text{Best\_}\Gamma \leftarrow \Gamma$ ;
    end if
     $n \leftarrow n - 1$ ;
  until ( $n = 0$ )
  return ( $\text{Best\_}\Gamma$ ); }

```

从上述算法描述中看出, 为避免标准贪心算法的问题, 回溯策略需维护每一步的当前状态. 而贪婪搜索当前状态每种分支情况可能需要比标准贪心算法更多的计算时间, 因每一步需要更多计算.

Algorithm 2 Greedy search for every direct branch of current state

```

/*  $\Gamma$ : current state of the partition structure */
/*  $\Delta Q_{ij}$ : the value by which  $Q$  increases if  $C_i$  and  $C_j$  are merged into a single community */
greedy_search( $\Gamma$ ) {
     $Q \leftarrow$  calculate  $Q_\Gamma$ ;
    do
         $\Delta Q^* \leftarrow -1$ ;  $i^* \leftarrow -1$ ;  $j^* \leftarrow -1$ ;
        for (each pair  $(C_i, C_j)$  within  $\Gamma$ ) do
            calculate  $\Delta Q_{ij} \leftarrow Q'_{ij} - Q$ ; /*  $Q'_{ij}$  is the  $Q$  value after merging  $C_i$  and  $C_j$  */
            if ( $\Delta Q_{ij} > \Delta Q^*$ ) then
                 $\Delta Q^* \leftarrow \Delta Q_{ij}$ ;
                 $i^* \leftarrow i$ ;  $j^* \leftarrow j$ ;
            end if
        end for
        if  $\Delta Q^* > 0$  then
             $\Gamma \leftarrow$  join_community ( $C_{i^*}, C_{j^*}$ );
             $Q = Q + \Delta Q^*$ ;
        end if
    until ( $\Delta Q^* < 0$ )
    return ( $Q$ ); }

```

5 仿真实验及性能评价

这部分将通过模拟仿真分析在第三部分提出的社区结构评价模型的性能。仿真程序中, 文档分布服从 Zipf 分布^[15]。每一个文档属于 ACM topic 的一个 topic 结点。这里在 Gnutella 拓扑的基础上模拟语义链接协议^[18]得到 SemreX 覆盖网络。语义覆盖网络的规模从 100 个结点到 1000 个结点。表 1 总结了语义覆盖仿真的相关设置。为了和 SemreX 系统设计保持一致, 这里将 Tsd 设置为 0.5, 因为系统测试表明当 Tsd 为 0.5 时系统能够取得较好的性能。

仿真实验中测试社区划分的值 Q , 如前面提到的, 达到最大的 Q 值时对应最优的社区结构。实验中, 网络规模为 1000, 分别计算出 Gnutella 拓扑网络和语义覆盖网的 Q 值。

图 5 显示了算法 1 计算的在不同社区划分情况下的 Q 值。图中语义覆盖网络被划分成 21 个社区时形成了最优的社区结构, 此时 Q 取得最大值 0.776。相比之下, Gnutella 网络表现了较差的社区结构属性, Gnutella 拓扑网络被划分成 85 个社区, Q 的最大值 0.570 明显比语义覆盖网的情况低。

图 6 显示结点数从 100 到 1000 不同规模情况下最大 Q 值的变化情况。图中显示, 语义覆盖网的社区结构的强度明显比 Gnutella 网络高。

图 7 比较 Gnutella 网络和语义覆盖网形成最优划分时的社区数。图中显示语义覆盖网的社区结构比 Gnutella 网络拥有的社区数明显较少。

以上实验结果表明语义覆盖网具有明显较强的社区结构的属性。接下来, 通过在语义覆盖网上模拟洪泛协议初步研究这种社区结构特征对系统性能的影响。

表 1 实验参数设置

Table 1 Settings for simulation semantic overlay

Parameter	Descriptions	Values
N	Number of nodes in the network	100~1000
T	Number of ACM topics	30
D	Max number of documents each node	200
Tsd	Threshold of similarity between peers	0.5
Q	Max number of queries each peer	200
$Z(\alpha, n)$	Zipf distribution of documents	$\alpha = 1.2$ $n = D$

图 5 不同社区划分情况下的 Q 值Figure 5 Optimizing Q in different communities division states in Algorithm 1 图 6 不同网络规模下 Gnutella 和语义覆盖网的 Q 值比较Figure 6 Comparison of maximum Q in Gnutella network and semantic overlay network 图 7 不同社区划分情况下的 Q 值

Figure 7 Comparison of number of communities of the best divisions in Gnutella network and semantic overlay network

图 8 不同网络规模下 Gnutella 和语义覆盖网的 Q 值比较

Figure 8 Comparison of recall rates in semantic overlay network and Gnutella network

图 9 不同网络规模下 Gnutella 和语义覆盖网的查询效率比较

Figure 9 Comparing efficiencies when varying the scale of nodes

这里主要用查全率 (recall) 来衡量仿真实验的效果, 它是信息检索中广泛采用的一个衡量标准, 是查询返回结果数占网络中所有相关文档数的比例, 表达如下

$$\text{Recall} = \frac{\text{Document}_{\text{relevant}} \cap \text{Document}_{\text{retrieved}}}{\text{Document}_{\text{relevant}}} \quad (12)$$

实验中同样采用规模为 1000 个结点的网络, 洪泛协议的 TTL 设置为 3. 图 8 表明语义覆盖网络的查全率比 Gnutella 网络好得多. 进一步将网络规模从 1000 扩展到 5000, 评估语义拓扑网络的搜索性能. 定义搜索效率为查全率和通信消息数之比, $\text{Efficiency} = \text{Recall} / \text{number of messages}$. 可见, 查询效率比单独看查全率或者通信开销更公平. 每个测试我们运行 20 次, 并报告结果平均值. 图 9 显示在所有网络规模下, 语义覆盖网络比 Gnutella 网络都具有更好的查询效率. 总体来说, 语义覆盖网络将查询效率提高了 71.4%.

基于以上实验得出结论, 充分利用社区结构的特性将提高 P2P 搜索效率.

6 相关工作

网络社区结构的研究成为目前复杂网络领域研究的一个新热点^[15,23], 最近的研究表明很多网络都拥有强大的社区结构特征, 包括 World Wide Web^[24]、文献引用网络、社会网络^[25]等等. 在文献[22]中, Newman 等人提出不带权网络的社区结构估算模型. 然而, Newman 的模型难以刻画语义覆盖网, 因为语义覆盖网络是一个复杂的带权网络^[26]. Newman 在他的模型求解中设计了标准贪心算法, 然而, 此算法在社区合并过程中不可逆, 难以逼近全局最优, 尤其是当起始选择不妥时将得不到好的结果.

本文的研究是首次尝试评价语义覆盖网中的社区结构. 这里对于语义覆盖网这种复杂的带权网络提出了新的社区结构评价模型, 并设计了回溯 - 贪心混合算法用于网络社区结构的划分.

7 总结和未来工作

本文研究了语义覆盖的一个重要特性——社区结构, 并通过 SemreX——一个用于语义 P2P 文献共享平台进行了具体的论述. 建立了社区结构特性的评价模型, 并设计回溯 - 贪心混合算法用于寻

找最优社区结构划分. 然后, 比较了 SemreX 语义覆盖网和 Gnutella 网络, 发现 SemreX 具有更明显的社区结构的特征. 此外, 基于 Gnutella 和 SemreX 两种覆盖模拟了网络洪泛协议, 结果表明具有明显社区结构特征的语义覆盖具有更好的搜索效率.

下一步, 将基于更大规模的语义覆盖网分析社区结构. 另外还将充分利用社区结构的特征设计更高效的搜索算法.

参考文献

- 1 Chen H, Jin H, Wang J, et al. Efficient multi-keyword search over P2P Web. In: Proceedings of WWW. 2008
- 2 Shen H T, Shu Y, Yu B. Efficient semantic-based content search in p2p network. IEEE Trans Knowl Data Eng, 2004, 16: 813–826
- 3 Stoica I, Morris R, Karger D, et al. Chord: a scalable peer-to-peer lookup service for internet application. In: Proceedings of ACM SIGCOMM'01. San Diego, 2001
- 4 Ratnasamy S, Francis P, Handley M, et al. A scalable content-addressable network. In: Proceedings of ACM SIGCOMM'01. San Diego, 2001
- 5 Li J, Loo B T, Hellerstein J M, et al. On the feasibility of peer-to-peer web indexing and search. In: Proceedings of IPTPS. Berkeley, 2003
- 6 Reynolds P, Vahdat A. Efficient peer-to-peer keyword searching. In: Proceedings of Middleware. Rio de Janeiro, 2003
- 7 Gnawali O D. A keyword-set search system for peer-to-peer networks. Master's Thesis. Massachusetts: Massachusetts Institute of Technology, 2002
- 8 Bender M, Michel S, Triantafillou P, et al. P2P content search: give the web back to the people. In: Proceedings of the 5th International Workshop on Peer-to-Peer Systems (IPTPS'06). Santa Barbara, 2006
- 9 Gkantsidis C, Mihail M, Saberi A. Random walks in peer-to-peer networks. In: Proceedings of IEEE INFOCOM'04. Hong Kong, 2004
- 10 Sripanidkulchai K, Maggs B, Zhang H. Efficient content location using interest-based locality in peer-to-peer systems. In: Proceedings of IEEE INFOCOM. San Francisco, 2003
- 11 Nejdl W, Wolf B, Qu C, et al. Edutella: a peer-to-peer networking infrastructure based on rdf. In: Proceedings of the 11th World Wide Web Conference (WWW'02). Hawaii, 2002
- 12 Nejdl W, Wolpers M, Siberski W, et al. Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. In: Proceedings of the 12th World Wide Web Conference (WWW'03). Budapest, 2003
- 13 Haase P, Broekstra J, Ehrig M, et al. Bibster: a semantic-based bibliographic peer-to-peer system. In: Proceedings of the 2004 International Semantic Web Conference (ISWC'04). Hiroshima, 2004
- 14 Li M, Lee W C, Sivasubramaniam A, et al. SSW: a small-world-based overlay for peer-to-peer search. IEEE Trans Paral Distr Syst, 2008, 19: 735–749
- 15 Iamnitchi A, Ripeanu M, Foster I. Small-world file-sharing communities. In: Proceedings of IEEE INFOCOM'04. Hong Kong, 2004
- 16 Newman M. Modularity and community structure in networks. Proc Nat Acad Sci, 2006, 103: 8577–8582
- 17 Jin H, Chen H, Ning X M. SemreX: a semantic peer-to-peer system for literature documents retrieval. In: Proceedings of the 1st Asian Semantic Web Conference (ASWC'06). Beijing, 2006
- 18 Jin H, Chen H. SemreX: efficient search in semantic overlay for literature retrieval. Future Gener Comput Syst, 2008, 24: 475–488
- 19 Ning X M, Jin H, Chen H. Efficient search for peer-to-peer information retrieval using semantic small world. In: Proceedings of the 15th International World Wide Web Conference (WWW'06). Edinburgh Scotland, 2006
- 20 Yu Y, Jin H. Building a semantic P2P scientific references sharing system with jxta. In: Proceedings of the 8th Asia Pacific Web Conference (APWEB'06). Harbin, 2006
- 21 Yuhua L, Bandar Z A, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng, 2003, 15: 871–882

- 22 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Phys Rev E*, 2004, 70, 0066111
- 23 Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. *Proc Nat Acad Sci*, 2004, 101: 2658–2663
- 24 Flake G W, Lawrence S R, Giles C L, et al. Self-organization and identification of web communities. *IEEE Comput*, 2002, 35: 66–71
- 25 Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discovering E-Communities. In: *Proceedings of the 15th international World Wide Web conference (WWW'06)*. Edinburgh Scotland, 2006
- 26 Barrat A, Barthélemy M, Pastor-Satorras R, et al. The architecture of complex weighted networks. *Proc Nat Acad Sci*, 2004, 101: 3747–3752

Finding and evaluating the community structure in semantic peer-to-peer overlay networks

CHEN HanHua & JIN Hai*

Services Computing Technology and System Laboratory, Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

*E-mail: hjin@hust.edu.cn

Abstract Recently much attention has been paid to semantic overlay networks for information retrieval in large scale peer-to-peer networks, and much research work on semantic overlay protocols and searching algorithms has been done and the results indicate that semantic overlay is efficient for content searching in peer-to-peer networks. However, very limited work has been done to analyze and evaluate the characteristics of semantic overlay networks. In this paper we identify a natural property of semantic overlay networks, the community structure. We propose a mathematical model to evaluate the property of community structure of semantic P2P overlay networks. A heuristic algorithm is designed to optimize the community structure. Using the evaluation model we compare the SemreX semantic overlay with the Gnutella network. Results demonstrate that a SemreX overlay network has the distinctive community structure feature, while a Gnutella-like network does not. We also simulate a simple flooding protocol in both overlays to show that the overlay with community structure is more efficient for content searching.

Keywords peer-to-peer network, community structure, semantic overlay, semantic similarity, search efficiency

computing and online social networks. He is a member of the IEEE.

CHEN HanHua was born in 1978. He received the Ph.D. degree in computer science and engineering from Huazhong University of Science and Technology in 2010, where he is now working as an Associate Professor. He worked at the Hong Kong University of Science and Technology as a research associate between 2009 and 2010, and as a visiting scholar between 2007 and 2009. His research interests include peer-to-peer

puting and grid computing, peer-to-peer computing, network storage, and network security. He is a senior member of the IEEE and a member of the ACM.

JIN Hai was born in 1966. He received the Ph.D. degree in computer engineering from Huazhong University of Science and Technology (HUST) in 1994. He is a Cheung Kung scholars chair Professor of computer science and engineering and the Dean of the School of Computer Science and Technology at HUST. He is the member of Grid Forum Steering Group (GFSG). His research interests include computer architecture, virtualization technology, cluster com-