基于密度估计的社会网络特征簇挖掘方法

韩毅[1]，方滨兴[1,2]，贾焰[1]，周斌[1]，韩伟红[1]
（1.国防科技大学 计算机学院，湖南 长沙 410073；2. 北京邮电大学，北京 100876）

摘 要：通过凝聚式聚类方法抽取网络的层次结构，并基于拓扑结构分析，给出了社会网络的标注密度估计函数。通过对密度估计函数在网络层次结构上的聚合操作，计算聚簇的特征性指标，从而达到发现特征聚簇的目的。在大规模的真实数据上对这些方法和模型进行了验证，实验结果表明，所提出的思路和模型是合理的，算法是高效、可伸缩的。

关键词：社会网络；特征簇；数据挖掘
中图分类号：TP311 **文献标识码**：A **文章编号**：1000-436X(2012)05-0038-11

Mining characteristic clusters: a density estimation approach

HAN Yi[1], FANG Bin-xing[1,2], JIA Yan[1], ZHOU Bin[1], HAN Wei-hong[1]
(1. School of Computer Science, National University of Defense Technology, Changsha 410073, China；
2. Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: A hierarchical structure extraction approach based on clustering was proposed, and a density estimation based on topological structure was designed. By conducting the hierarchical aggregation on layers of hierarchical structure, the characteristic of clusters could be measured. The empirical study conducted on a large real data set indicates that the model and measures are interesting and meaningful, and the algorithms are effective and efficient in practice.

Key words: social network; characteristic clusters; data mining

1 引言

近年来，社会网络在研究上获得了广泛的关注，主要研究包括节点的度分布分析[1]、个体排名[2,3]、模式挖掘[4]等。实际中，许多社会网络，例如在线交友、论坛等，都体现出了极强的社团化特征。在当前的社区识别方法中，通过边密度识别社区是一种直观而常见的方法，即对于一组节点，如果节点间的内联密度与外联密度的差值满足阈值差值，即可将这组节点视为一组社区。基于边密度的方法是一种单纯的拓扑结构分析方法，通常不考虑节点的属性信息。而大多数社会网络都是标注网络，即节点上往往蕴含着大量的属性信息，基于拓扑结构的社区发现方法往往缺乏对节点属性特征的解释能力[5]。

社会网络中的社区可能是真实社会关系的映射，例如，在在线交友网站中，一个社区可能是具有某种共同兴趣爱好的网络群体；在学术合作网络中，一个社区可能是共同研究方向的一群学者等。因此，分析社区中链接紧密的群体相互作用的原因，解释社区的形成机理就显得非常重要。在实际生活中，节点上承载的信息往往会作用于网络，促

收稿日期：2010-12-27；**修回日期**：2011-05-25
基金项目：国家自然科学基金资助项目(60933005，60873204)；国家高技术研究发展计划（“863”计划）基金资助项目(2010AA012505)
Foundation Items: The National Natural Science Foundation of China (60933005, 60873204); The National High Technology Research and Development Program of China (863 Program) (2010AA012505)

进网络结构的演化；而网络结构往往又会反作用于话题，促使话题影响的放大。举例来说，在线社交网站中围绕某个热点话题讨论时，网络常常是某些话题的催化剂，话题影响力往往可以随着某个群体的参与而被迅速升温，而话题的讨论，往往会促使陌生节点通过话题产生关联，从而促成了新的社会关系。在互联网上，当一个话题迅速爆发时，如何判断其是由于人们本身对话题的兴趣使然，还是网络结构的放大作用导致了这种情况的发生？一个热点的崛起是否有人为的背后推手？在这种背景下，判断网络的链接结构和其承载话题的关联就显得意义重大。研究这种网络链接关系与话题现象之间的相互联系，在交友社会关系推荐，广告精确投放，舆论情报挖掘等应用领域，都具有重要的理论和实践意义。

当判断话题和网络结构的内在联系时，首先独立分析链接关系和文本标注，再进行线性组合，是一种直观的解决方案，然而，常见类似 $ax_{\text{link}} + ßx_{\text{topic}}$ 链接+话题的加权模型，其实际含义往往很难解释，而且参数设置方法也不确定[5]；在实际操作中，基于链接发现的聚簇和基于标注信息发现的聚簇并不能完全对应，算法产生的中间结果多，运算开销也不能很好地满足实际应用的需求。

针对上述问题，考虑将社会网络中的链接关系分析和文本特征挖掘这 2 个方面统一起来，设计了一种特征聚簇挖掘方法。特征簇是网络中高链接密度的节点簇，其标注信息在分布上也具有高密度、代表性的特点。本文给出一种特征簇发现方法，即通过分析网络的链接结构，将网络的链接密度分布层次提炼出来；在抽象链接层次结构的同时，对标注信息进行密度估计和特征指标的计算，通过估计特征指标的上下界，对结果集搜索进行预先剪枝，从而达到快速挖掘特征簇的目标。

本文的主要贡献包括：首先，在分析了社会网络聚类和社区发现相关技术的基础上，设计了基于凝聚式聚类的社会网络层次结构提取方法，将在扭曲空间中的社会网络结构，通过提取层次式提取，映射到空间曲面上来，形成一种类似于等高线地形图的密度层次结构；其次，给出了一种基于网络链接结构的标注密度估计方法，即将网络上离散的标注信息，通过一种类似于疾病传播原理的模型，抽象为标注密度函数，并通过在层次结构中不同粒度簇中的聚合计算结果，估计标注在链接结构中各处的特征指标；再次，设计了特征簇的挖掘算法，即在抽取网络层次结构的同时，通过估计各部分的特征函数的上下界，设计了自底向上的剪枝策略，实现高效精确计算特征簇的算法；最后，在大量真实的数据集上的实验验证了模型和算法，结果证明提出的算法是具现实意义的，算法是高效的。

2　相关工作

本文的工作和社会网络上的聚类分析、社区识别、模式发现等研究工作具有紧密联系，本节简要回顾这些技术。

在目前的聚类或社区发现技术中，基于图结构约束进行社区发现是一种主要的技术手段，文献[6]和文献[7]使用了图论中的完全子图和准完全子图定义社区的基本结构，伊利诺伊厄巴纳香槟分校的 Karypis 等人[8]在准完全子图的基础上进行扩展，定义了一种边密度约束的社区定义方法；由于最大完全子图的发现早在 1972 年已经被证明为是 NP 难问题[9]，上述工作都采用了贪婪算法，并且针对完全子图对节点度的约束条件进行剪枝；在经典图论中，节点间的可达性[10]、介数[11]、边密度[12]等属性也都可以作为约束条件，从而对节点进行聚类或社区划分。NEC 普林斯顿研究院的 Flake 在研究 Web 链接结构的同时，提出了使用最大流—最小割定理[13]对其进行社区进行划分的方法，其基本思想是将网络模型化为信息流通的信道和关节，根据其边路的信息通过能力判断其社区边界。密歇根大学的 Newman 等提出了模块性的概念[14]，认为社区之所以有别于其他网络结构，在于其内联边密度应该明显大于一般随机网络的边密度期望，通过学习网络全局属性从而自动阈值，并通过节点的反复组合从而优化节点群的模块性。国内，中科院计算所的程学旗等提出了使用信息瓶颈模型来发现社区[15]，通过寻找网络的最优压缩表示来发现网络的社区结构。

社会网络上的特征模式发现近年来也受到了广泛的关注，基于链接结构的频繁模式挖掘是其中一个主要方向，其主要思想是，给定一个支持度阈值，在网络中发现频繁程度不低于这个阈值的频繁子结构；在这种定义下，人们通常使用能否符合给定的同构映射来判断 2 个子结构是否相等。与在事务性数据上的频繁模式挖掘算法类似，

主流的频繁子结构挖掘算法也可以概括为基于Apriori算法的方法[16~18]和基于Pattern-Growth的方法[4];基于Apriori的挖掘方法需要生成图模式的候选集，与传统事务性数据库不同，图模式增长不但需要考虑节点的扩展，还需要同时考虑边的扩展，带来的组合爆炸问题非常明显，并不适用于大规模的网络数据；伊利诺伊香槟分校的Yan、Han等人在2002年提出了一种基于Pattern-Growth的频繁模式挖掘算法gSpan[4]，为了避免发现重复的结构，他们给出了一种右路优先的遍历策略；与之类似的方法还有FFSM[19]、SPIN[20]等。上述的频繁特征抽取方法可以有效抽取网络中的结构特征，但是都没有考虑节点属性相关的模式，而且基于同构子图的特征模式定义也并不适用于发现规模较大的模式特征。随着信息检索技术和搜索引擎技术的进步，基于节点属性/主题进行的社会网络模式发现也受到了更多的关注，香港中文大学的程红等人[5]设计了一种混合的标注图划分方法，通过自适应的识别节点链接结构和标注信息的相似性，达到混合聚类的目标。该工作虽然将二者有机统一，但在全局视角上，其识别的聚簇并不能保证内联紧密，而且也不能满足挖掘特征簇的要求。

3 网络结构层次提取

现实生活中，许多社会网络，例如在线交友网站、学术合作网络、通信网络、生物蛋白质相互作用网络，都可以被模型化为图。

在本文中，一个社会网络由三元组$G=\langle V,E,L\rangle$表示，其中，$v\in V$表示个体和个体集，个体间的关系由向边$e=\{u,v\}\in E$表示，E代表边（链接）集，L是节点上的标注函数，$l\in L(v)$表示标注l是节点v上的关键词之一。

一个簇$C\subseteq G,V$表示社会网络上的一组节点，对于任意一个标注l，簇的特征指标是二元组$\langle l,C\rangle$的函数，表示为$spc_l(C)$，其取值为标注l在簇C上的显著性。给定一个参数n，特征簇挖掘就是找到使$spc_l(C)$取值最高的n个二元组。

其现实意义就是，在在线社交网站中，看哪些人专注于讨论哪些事；在学术合作网络中，看哪些团队专注研究哪些方向等。

传统基于边密度的社会网络聚类技术往往需要设置阈值用于控制簇内聚和外联的比例关系，或者通过自适应的方式隐式地设置类似的参数。然而，在真实社会中，社区间并不存在绝对的划分边界。因此，希望能够设计一种方法，可以将扭曲空间上的网络结构映射为链接密度的分布函数，从而可以在各个层次上分析链接密度和其标注信息的关联。如图1所示，图中阴影强弱表示对应链接密度高低。

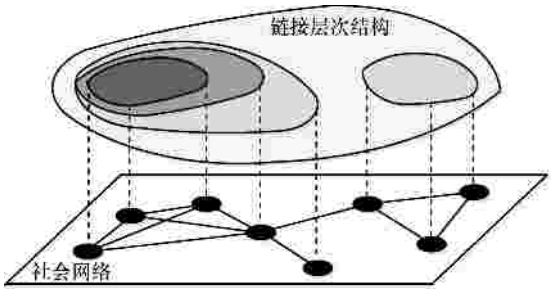


图1 链接密度结构

3.1 基于凝聚式聚类的层次结构提取

本文使用了一种基于凝聚式聚类的方法，将聚簇的过程理解成一种节点间自我组合的过程。在最开始的时候，每个节点都隶属于一个单独的簇；每次迭代将2个“距离”最近的簇聚类在一起；簇间的“距离”由其边密度来定义，即2个簇合并后形成的边密度越高，这2个簇就越接近。当所有的节点都被聚类成为一个单独的簇时，这个过程终止。与一般凝聚式聚类不同的是，并不要求设置凝聚过程的终止条件，节点的凝聚过程一直持续进行，直到网络中全部个体都聚簇在一个单一的簇中。这一过程中，利用节点间相互聚簇的中间结果表示网络的层次结构。

这个凝聚过程中的层次结构(dendrogram)，记为D，可以被理解为是这个社会网络链接紧密程度的表示。形式化来说，一个社会网络$G=\langle V,E,L\rangle$的层次结构是一个簇的集合$D_G=\{C_1,\mathrm{L},C_{2|G,V|-1}\}$。它是一个二叉树。$D_G$中的每个节点$C\in D_G$都代表节点$v\in G,V$的一个聚簇，即$C\subseteq 2^{G,V}$（$2^{G,V}$表示节点集$G,V$的幂集），其中，$D_G$中共有$G,V$个叶子节点簇，即每个$v\in G,V$都独立成簇，即$\forall v\in G,V$，都有$\{c\}\in D_G$，记为$C_{\text{leaf}}$；$D_G$中还有$|G,V|-1$个非叶节点簇，每一个非叶节点簇$C_{\text{nonleaf}}$都是由其后继簇通过合并操作而来，可以理解为是凝聚式聚类的一个中间结果，即$\forall C_{\text{nonleaf}}\in D_G$，$C_{\text{nonleaf}}=\langle C_{\text{left}},C_{\text{right}}\rangle$，其中，$C_{\text{left}},C_{\text{right}}\in D_G$。

如图 2 所示为 6 节点的一个社会网络的凝聚式聚类层次结构，其后续遍历次序重现了层次式聚类的过程。对于$|G,V|$个节点的社会网络，每次迭代合并 2 个距离最近的社区，共计需要$|G,V|-1$ 次合并。算法 1 描述了凝聚式聚类和聚合层次的抽取方法。

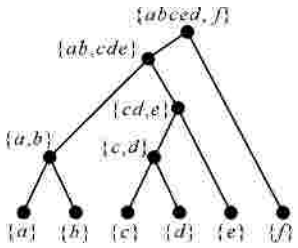


图 2　网络链接密度层次结构的例子

算法 1　凝聚式聚类和层次结构提取

输入：图$G=\langle V,E,L\rangle$**；**

输出：层次结构图D_G**；**

让T **为临时空间；**

让$T=D_G=\{\{v_1\},\cdots,\{v_n\},v_i\in G,V$ **；**

While $|T|>1$

　随机挑选$C\in D$ **；**

　找到一个$C'\in D$**，使**$C'=\arg\max(r(C\cup C'))$ **；**

　$T=T\cup\{C'\cup C\}-C'-C$ **；**

　将~$\{C'\cup C\}$~**加入**~D_G **；**

End While

返回D_G **；**

由于凝聚式聚类方法每次仅选取 2 个距离最近的簇进行合并，现实中往往会存在 $n(n>2)$个簇间两两距离最近且相等的情况。如果遇到 $n(n>2)$个簇间两两距离最近且相等的情况，一种解决方式是在 n 个簇中随机地挑选 2 个先合并，之后再逐个合并其他；或者将 n 个簇同时合并，则层次结构图中将会出现带有 $n(n>2)$个后继的节点，簇总数会小于 $2|G,V|-1$；前者的节点簇集合是后者的超集。由于对层次树进行后续遍历可以视为是对聚类过程的重现，所以尽管 2 种方式生成的簇总数不同，但可以很容易看出，其后续遍历的次序是相同的。因此这 2 种方案的结果是相容的，且很容易相互转换。由于二叉树的方案可以生成更多的簇，有利于更细粒度表示层次结构，因此本文采用了二叉树的解决方案。

3.2　基于边密度的簇间距离

凝聚式聚类的关键在于确定最近的簇，即定义簇间距离。在本文中，采用了基于边密度的簇距离衡量方法。如果将 2 个簇组合，新簇的边密度越高，表示这 2 个簇的距离就越近。需要注意的是，簇间距离的定义并不是唯一的，任何符合凝聚式聚类要求的距离定义方式，都可以应用到这里，例如，最大流最小割定理[13]的距离定义，或采用信息瓶颈方式的定义[15]等。

定义 1　**边密度：对于图**$G=\langle V,E,L\rangle$**和节点簇**$C=v_1,\cdots,v_n\subset G,V$，$C$ **的边密度定义为**

$$r(C)=\frac{2|E_{C\subset G,V}|}{|C|(|C|-1)}$$

其中，$E_{C\subset G,V}$ **表示簇**C**在图**G**中的导出子图**(induced subgraph)**的边集。**

对于一个 n **节点簇，至多有**$\binom{n}{2}$**条无向边。**

$r(C)$ **的现实意义是**$|C,E|$**在最大可能边数中的比重(无向图，每** 2 **个节点间仅允许一条边)，因此，有**$0\le r(C)\le 1$**。**

定义 2　**簇的最近邻集合：对于一个图**$G=\langle V,E,L\rangle$**和其上的导出子图集合**$W=\{C_1,\cdots,C_n\}$**，对于任意簇**$C\in W$，C **在**W**中的最近邻，定义为**

$$NN(C)=\{C'\in W\mid \text{ó}C''\in W: r(C'',C)>r(C',C)\}$$

簇C**的最近邻是一个簇集合，其含义是在**W**中，无法找到** $NN(C)$**以外的其他簇**C''**，与** C **合并可以获得更好的边密度。凝聚式聚类的主要思想就是每次从** W **中随机选择一个簇** C**，并在** C **的最近邻集合** $NN(C)$ **中一个簇与** C **合并，并将合并的新簇置入** W **中。**

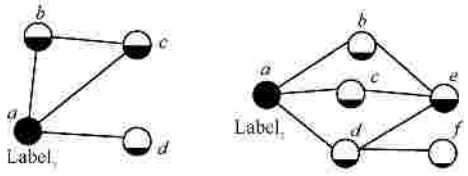
4　标注密度估计

一个静态的社会网络数据集 G 可以被视为是一个动态演化网络的快照，而静态网络上的标注 L 则是标注的采样，标注密度估计的目的就是还原这种标注的分布函数。与可度量空间上的核密度估计类似，标注 $l\in L(v)$ 被视为标注密度函数采样的一个样本，其密度分布函数应在 v 处的取值应最高，其他节点的密度函数取值应以标注所在节点为中心，沿网络路径指数级衰减。

本文采用了传播模型(PRM, propagation model)的方式估计标注密度。在当前流行的社交网络，例如人人网、新浪微博、Twitter 等，都可以被模型化为图$G=\langle V,E,L\rangle$，$v\in G,V$ 表示社交网络的 ID 集，E 表示社交网络中的关注/好友等社会关系，L 表示

为带有相应 ID 发表的内容信息。当某个 ID 发表带有一定主题的信息后，认为其相应的社会关系会在一定概率下阅读并继续传播该信息。这个传播模型与病理学中疾病的传播模型类似，即将带有某个标注 l 的节点 v 视为感染源，其影响会依据一定的概率传播到其相邻的节点上，当一个节点的较多邻居都被 l 感染，此时该节点感染 l 的概率也会增加。标注信息的影响力会随着网络路径的延长而受到指数级衰减，而衰减的程度由用户指定的阻尼因子来控制。

如图 3 所示，Label_1 为节点 a 所带标注，图中节点阴影部分表示标注 Label_1 在其他节点的传染情况。如图 3(a)中，Label_1 的权值随 a 的链接经过衰减传染到其相邻节点 b、c、d 上。图 3(b)中举例说明了 Label_1 的影响力经过 2 级传播的情况。值得注意的是，仍然是在静态网络的前提下，讨论通过估计标注信息的传播模型来对标注密度进行估计，该模型可以很容易扩展到演化网络上，受篇幅所限，本文不分析演化网络的情况。



(a) 标注密度传播(1 级) (b) 标注密度传播(2 级)

图 3 标注密度估计的例子

与核密度估计类似，一个标注在网络各点的标注密度值总和应与其出现次数一致。假设一个节点 v 上带有标注 l，即 $l \in L(v)$，记为 l 的一次出现。设 l 在 v 上的密度函数取值为 w，l 会扩散到其相邻节点上，其权值也会相应衰减为 $l\,w$。为了公平起见，约定 l 的一次出现对整个社会网络的贡献为 1。换句话说，如果在 G,V 上，l 仅在一个节点上出现过一次，则所有节点上关于 l 的密度之和为 1，如果 l 出现了 n 次，则 l 在所有节点上的密度总和为 n。

定义 3 标注密度估计 对于图 $G=\langle V,E,L\rangle$ 和节点 $u,v\in G,V$，如果有 $l\in L(v)$，则 l 在 u 上的出现对 v 的影响为

$$imp_l(u\to v)=\begin{cases}(1-l)\sum\limits_{p\in P_{u\to v}}\dfrac{l^{|p|}}{d(p_i)}, & u\neq v\\ (1-l), & u=v\end{cases}$$

表示 l 在 u 上出现对 v 的效应累积。其中，$P_{u\to v}$ 表示所有从 u 到 v 的无环路径集合，路径 $p=\langle p_0,\text{…},p_n\rangle\in P_{u\to v}$，其中，$u=p_0, v=p_n$ 表示路径两端，$|p|=n$ 为路径长度，$d(p_i)$ 为节点 p_i 出度，$0<l<1$，为阻尼因子。

节点 u 上标注 l 的密度为

$$den(l,u)=\sum_{v\in Ins(l)} imp_l(v\to u)$$

表示 l 所有出现对 u 的影响的总和，其中 $Ins(l)=\{u\in G,V\mid l\in L(u)\}$ 表示 l 的出现集，$|Ins(l)|$ 表示标注 l 在 G 中总计出现次数。

对于一个簇 C，标注 l 在簇 C 上的密度为

$$den(l,C)=\frac{\sum_{v\in C} den(l,v)}{|C|}$$

表示 l 在 G 的所有成员上的密度均值。

若有 $l\in L(u)$，如定义 3 描述，$v(v\neq u)$ 获得的来自 u 的影响与 $u\to v$ 的距离成反比。设想将图以标注所在节点为中心展开，图 4 举例说明了图 2(b)中的依据路径展开的情况。分配到节点 e 的密度值 $imp_{\text{Label}_1}(a\to e)$ 来自由 a 通往 e 的 3 条路径，f 与 e'' 共享来自第 3 条路径的密度。e、e'、e'' 获得密度的总和即为 $imp_l(a\to e)$。

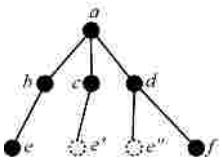


图 4 图 3(b)所示例子的路径展开

当 $l\in L(u)$ 时，设 $D_i(u)$ 为与 u 距离为 i 的所有节点的集合，可以很容易看出，分配给 $D_i(u)(i=0,1,\text{…},n)$ 的密度值分别为 $den(1,u), l\,den(l,u),\text{…}, l^n den(l,u)$，当网络无限大时，密度的总和为 $\dfrac{den(l,u)}{l-l}$。由于 l 一次出现在网络上的密度贡献应为 1，因此 $den(l,u)=1-l$。算法 2 给出了网络上节点标注密度的计算方法。

算法 2 节点标注密度计算方法

输入：图 $G=\langle V,E,L\rangle$，参数 l；

输出：$den(l,v)$；

初始化密度矩阵 $\boldsymbol{G}=j_{|v|\times|L|}=0$；

FOR $\forall l\in L(v)$

IF (v 已被访问 $d(v)$次)

跳出；

ENDIF
FOR $\forall l \in L(v)$
调用 traverse(v，l，1- l)；
ENDFOR
ENDFOR
返回 $den(l,v)=\boldsymbol{G}$ ；
算法 3　递归子过程：*traverse*()
输入：v，l，*value*；
让 $j_{v,l}=value$ ；
IF v 已被访问 $d(v)$次
跳出；
ENDIF
FOR 对于 v 的每个邻居 u
调用 traverse($u,l,\frac{l\ value}{d(v)}$)；
ENDFOR

5　挖掘特征簇

在第 4 节讨论了通过凝聚式聚类提取社会网络层次结构的方法。层次结构图中，每个节点表示一个凝聚式聚类的中间结果，即一个簇。节点上的标注密度函数表示了该位置相应关键字所对应的影响力强度。

5.1　特征指标

当某个标注在某个簇内的分布越密集，而在其他部分的出现约稀疏，则可以认为这个簇的特征越显著。

定义 4　l-标注显著性：对于图 $G=\langle V,E,L\rangle$ 和簇 $C\subset G,V$ ，给定一个标注 l，簇 C 的 l-标注显著性，表示为

$$spc_l(C)=\frac{P(l\mid C)}{P(l\mid C^C)}=\frac{den(l,C)}{den(l,C^C)}$$

其中，$C^C=\overline{C}$ 表示 C 的补集(complementary set)。

对于一个节点簇 C，其 l-标注显著性是 l 在 C 簇内密度与簇外密度的比值。而本文的目标就是挖掘具有大标注显著性的节点簇。

5.2　特征簇挖掘方法

特征簇可以理解为社会网络中具有强烈标注特征性的簇，即给定一个参数 n，对于任意标注 l 和层次结构中的任意簇 C，使 $spc_l(C)$ 最大的 n 个簇和标注的组合。

最直观的方法计算 top-n 特征簇，首先，抽取社会网络的层次结构；其次，对于每个标注，通过计算其在网络上的密度分布；对于每个标注和簇的二元组 $\langle l,C\rangle$ 计算其特征性指标，并使用阈值算法(threshold algorithm)输出 top-n 的结果集。

上述的方法需要计算所有标注和簇的组合，计算开销较大，因此，可以通过以下策略进行剪枝。一个簇的特征指标可以看作是其所有子簇的特征指标根据子簇大小进行加权平均的结果。

定理 1　特征指标边界 对于图 $G=\langle V,E,L\rangle$ 和 2 个簇 $C_1,C_2\subset G,V$ ，对于任意标注 l，有

$$\min(spc_l(C_1),spc_l(C_2))\le spc_l(C_1\cup C_2)\le\max(spc_l(C_1),spc_l(C_2))$$

因此，一个大簇的特征指标的上下限可以通过定理 1 来估计，即给定一个特征指标的阈值 ?，对于层次结构图中的一个节点，若其所有后继的特征指标都不超过 ?，那么其本身的特征指标也不会超过 ?。

定理 2　密度边界：对于图 $G=\langle V,E,L\rangle$ 和簇 $C\subset G,V$ ，对于任意标注 l，有

$$den(l,C)\le\frac{|\{v\in G,V\mid l\in L(v)\}|}{|C|}$$

$|\{v\in G,V\mid l\in L(v)\}|$ 表示标注 l 在图 $G=\langle V,E,L\rangle$ 中出现的次数，等于 l 在所有节点上密度分布的总和。显然 l 在簇 C 上的特征指标不可能超过 $\frac{|\{v\in G,V\mid l\in L(v)\}|}{|C|}$。

因此，设计了一种将层次抽取，密度估计和特征指标计算综合在一起的计算 top-n 特征簇的方法。

其主要思想是，使用自底向上的方式生成层次结构图，每生成一个簇，同时估计节点标注在该簇上的密度分布和特征指标的上下界，通过不断更新进入 top-n 簇的特征阈值，提前确立或淘汰候选。算法 4 给出了伪代码。

算法 4　抽取特征指标最大的 n 个二元组 $\langle l,C\rangle$
输入：图 $G=\langle V,E,L\rangle$ ，参数 n；
输出 特征指标取值最大的 top-n 二元组 $\langle l,C\rangle$ ；
初始化结果集 $C=f$ ；
初始化 top-n 阈值 $v=0$ ；
按照算法 1 的方法获得一个簇 C。
WHILE C 不是层次图中的根节点
FOR $\forall l\in L(C)$
根据定理 1 和定理 2 的方法估计 $\langle l,C\rangle$

的上下界；

IF{ $\langle l,C\rangle$ 的上界小于 v }

跳过；

ELSE

计算 $\langle l,C\rangle$ 的特征指标 $spc_l(C)$，并更新 v 和 C ；

ENDIF

ENDFOR

ENDWHILE

返回 C ；

6 实验分析

在真实的社会网络数据集上验证了本文的模型和算法的正确性，并在一个大规模的数据集上测试了算法的效率。

算法使用 Java 的实现，具体软件和硬件环境如表 1 所示。

表 1 实验环境

CPU	Intel Core 2 Q9550 2.83GHz
内存	8GB
操作系统	CentOS 内核 2.6.18
运行环境	Java Runtime Environment 1.6

6.1 数据集

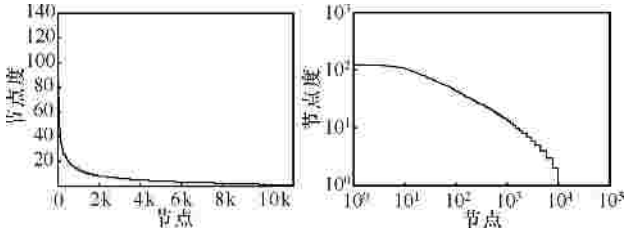
本文挑选了数据库相关的 9 个国际会议(SIGMOD、VLDB、PODS、ICDE、ICDT、DOOD、EDBT、SSD 和 CIKM) 从收录起到 2010 年 4 月的论文作为基本素材。

这个数据集包含 12 004 个作者，10 372 篇论文及其 34 173 个合作关系(双向边)。数据可以在 DBLP(\url{http：//dblp.uni-trier.de/xml/})下载。

在该数据集中，每个节点表示一个作者，如果 2 个作者之间合作过论文，那么他们之间就会存在一双向条边，作者发表论文的标题信息(英文)则作为节点的标注，称这个数据集为 DB 数据集。

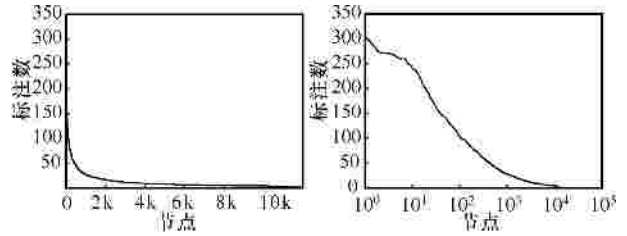
对于节点上的标注信息，为了避免英文词型变化对标注信息的影响，使用了波特词干算法(porter stemming algorithm)[21]对标注进行了处理，并使用了一个大小为 150 词的非检索用字表(stopword list)。

从图 5~图 7 可以看出，这 2 个数据集在节点度分布上都呈长尾形式，从对数刻度图分析，其分布接近幂率分布。



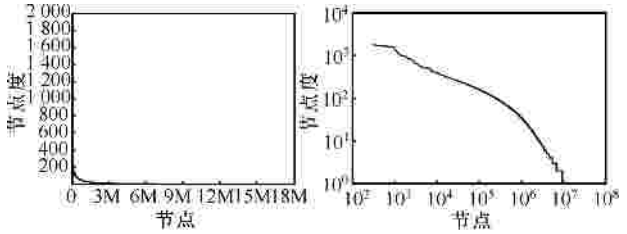
(a)标准刻度 (b)对数刻度

图 5 节点度分布：DB 数据集



(a)标准刻度 (b)对数刻度

图 6 标注信息分布：DB 数据集



(a)标准刻度 (b)对数刻度

图 7 节点度分布：twitter 数据集

从节点数/链接数比可以看出，这 2 个社会网络都属于稀疏图，即 $|E| \ll \binom{|V|}{2}$。

此外，于 2010 年 4 月在境外 twitter (\url{http：//snap.stanford.edu/data/twitter7.html}) 网站上通过其官方 API 抓取了公开 ID 的一个连接构件 (connected component)，每个节点表示一个注册 ID，其 ID 间的 WFW (who follows whom，twitter 中的关注)关系构成有向边集，共截取了 ID 在 120h 内的发表内容作为 ID 的标注，共计包含 18 012 823 个注册 ID 和其对应的 33 237 699 个 WFW 关系，本文中称该数据集为 TW 数据集，将该数据集根据空间和时间分别进行切片，用于测试本文给出的方法的加速比。

6.2 凝聚式聚类过程分析

图 8 展示了算法在使用凝聚式聚类生成层次结构图时，每次凝聚合并时，生成的新簇的节点数与合并顺序的关系，其中簇大小为对数刻度，DB 数据集。

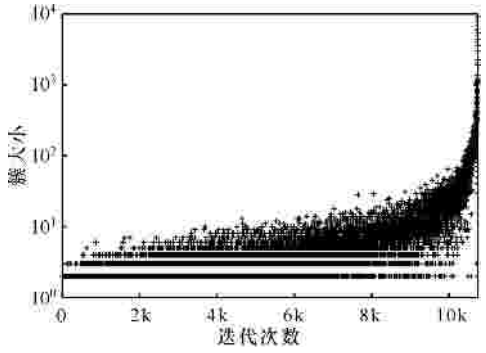
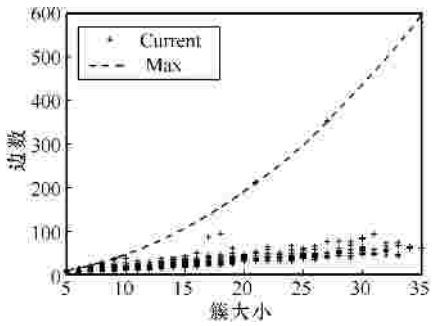


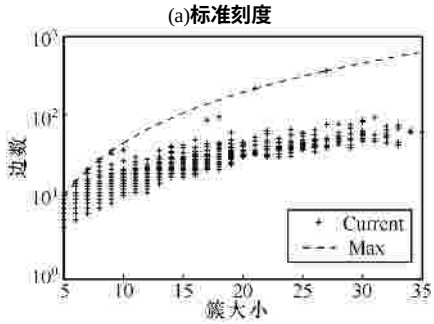
图 8　每次迭代合并簇的情况

可以看出，在最开始的时候，簇的尺寸较小，随着凝聚式聚类的延续，簇的节点数不断增加。总体来看，簇的合并在宏观上呈一个平稳的态势，这个例子说明本文的凝聚式聚类过程倾向于将节点数较少的簇先合并，而不是将小簇向大簇靠拢。这样的合并方式有利于生成较矮的层次结构，链接密度的等高线层次少，坡度平缓，这也从一个侧面说明，本文使用的 DB 数据集中，大部分小簇之间的位置相对平等。

图 9 展示了凝聚式聚类后，层次结构中簇内的链接数和簇大小的关系，每个数据点代表一个簇，实线是 $y=\binom{x}{2}$ 的曲线，表示一个簇理论上可以具有最多的双向边的数量。为了表示方便，图中仅标出了 5 个节点以上簇的情况。这个图说明生成的层次结构图中每个簇的内聚很高，远远高于数据集的平均链接密度，层次式聚类的效果是明显的。



(a)标准刻度



(b)对数刻度

图 9　簇密度分布：DB 数据集

6.3　结果分析

本文输出了社区成员数大于 3 且标注显著性指标得分最高的 3 000 个二元组 $\langle l,C\rangle$ 的情况，图 10 中，每个数据点表示一个簇，对应的横坐标表示其特征指标得分，纵坐标表示对应的簇大小，设置阻尼因子 $l=0.5$。

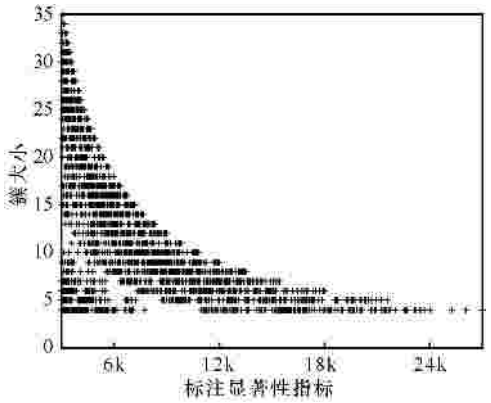


图 10　标注显著性指标——簇大小分布：DB 数据集

从图 10 可以看出，得分较高的簇尺寸小，得分较低的簇尺寸大。由于采用了“人均密度”的概念定义标注显著性指标，因此，上述结论符合常识。然而，在图 10 中可以看出一个有趣的现象，即在数据分布中存在一个明显的断层，这是因为在数据集节点的标注信息中，往往存在部分高频的专有名词或缩写，这部分标注分布集中，因此支配结果集的上层部分。本文认为，这也可以作为从数据集中提炼专有名词的一种手段，可以应用于今后的研究中。

由于本文中提出的特征挖掘是在社会网络的层次结构树上展开的，因此，在运算中不可避免地会遇到簇间重叠和覆盖问题，即在自底向上的挖掘过程中，子簇和其父簇同时出现在结果集中的情况。在表 2 中分析了图 10 中 3 000 个簇的重复情况。在 3 000 个得分最高的簇中，超过 70% 的簇间重叠次数小于等于 3，结果集冗余程度较低。

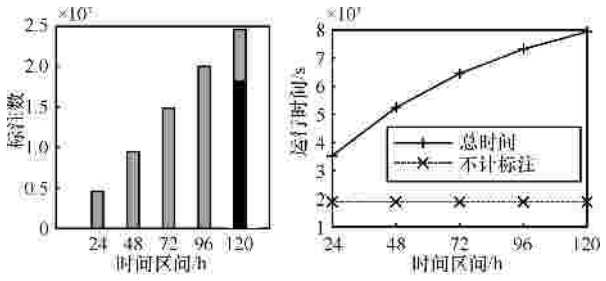
6.4　效率分析

DB 数据集的规模较小，因此，在实现时将整个链接结构可以完全载入主存中进行计算。为了避免系统虚存对算法效率的影响，使用 TW 数据集对算法的伸缩性进行测试。

表 2 标注显著性指标得分最高的 3 000 个簇的簇间重复情况

重复次数	数量	比例/%
10	3	0.1
9	3	0.1
8	11	0.37
7	64	2.13
6	133	4.43
5	218	7.27
4	359	11.97
3	602	20.07
2	944	31.47
1	663	22.1

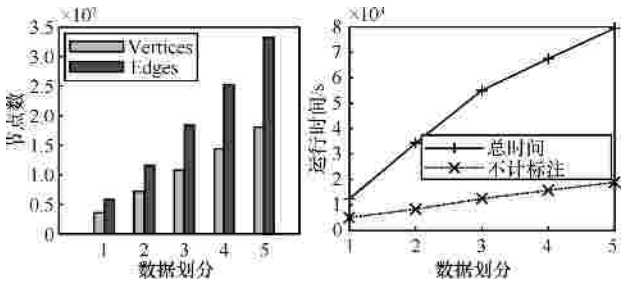
首先，测试标注数量和运行时间的关系。累积采集了 120h 的 twitter 用户发言，将标注与网络拓扑结构分离，对标注数据以 24h 为一单位进行分片，分别计算 24h、48h 等数据量下算法的运行时间。图 11(a)所示每 24h 的标注数量分布，可以看出，标注数量随时间变化呈线性增长。图 11(b)展示了算法的运行时间，宏观上看，算法的运行时间接近于线性时间。为了更清楚地表示算法每一部分所耗费的时间，还给出了仅进行凝聚式聚类(去掉标注信息)所需要的时间作为参考。由于标注数量随时间线性增长，因此进行标注密度估计的时间也随之增加，但是由于采用了剪枝策略和阈值算法，时间曲线的末端明显低于线性时间。



(a)标注数 (b)运行时间

图 11 标注数与运行时间

为了测试节点数量和运行时间的关系，首先使用随机游走的方法对网络进行采样，递增式地取了 5 个不同规模的子数据集，用于测试算法在不同节点/链接规模下的伸缩性。图 12(a)展示了每个部分节点数和边数的分布情况，图 12(b)展示了总运行时间和去除标注后的运行时间。由于节点规模较大，内存中不能完全存储社会网络的邻接表。社会网络上的凝聚式聚类与可度量空间上的聚类不同，社会网络上的最近邻计算的候选仅仅需要在相邻的簇中选择即可，而从本节的节点出度分布可以看出，绝大部分节点相邻节点数都在同一水平上，因此，凝聚式聚类所需要的时间接近线性，是可以理解的。由于层次结构树中簇的数量和节点数在同一量级，因而进行凝聚式聚类并生成层次结构的时间也近似于线性时间。随着节点数量的增加，标注信息也在线性增加，因此，总运行时间也接近于线性时间。



(a)节点数 (b)运行时间

图 12 节点数与运行时间

综合分析原因，由于社会网络中凝聚式聚类的特殊性，链接层次结构提取过程并不占用很大的运行时间，且接近于线性时间。因为 TW 数据带有大量标注信息，运行时间的大部分被标注数量统计和密度估计部分支配。

目前，还没有其他社会网络特征挖掘算法考虑在层次式聚类的基础上进行标注密度估计和特征提取，但是有部分综合考虑属性和结构 2 个要素的社会网络节点聚类算法可以达到类似的效果。将本文提出的方法与近年在国际知名数据库会议提出的相关算法在 twitter 子集的运行时间进行了比较。SA-Cluster[5]是一种综合考虑节点属性和结构特性社会网络节点聚类方法，通过自适应的识别节点链接结构和标注信息的相似性，达到混合聚类的目标；Inc-Cluster[22]SA-Cluster 的改进算法，通过增量式更新提高算法效率。与数据挖掘中的 k-means 算法类似，上述算法均需预先设定目标簇数 k，而上述算法效率都对 k 的设置较为敏感，因此，分别测试了 k=10^4、10^5、10^6 的情况。tfidf[23]是一种仅考虑社会网络节点属性，通过词频统计进行特征计算的算法。

通过图 13 所示的结果可以看出，由于 tfidf 并未考虑社会网络上节点的结构属性，因此其获得了很高的执行效率。本文所示算法比 Inc-Cluster 算法在 $k=10^4$ 时略优。由于 SA-Cluster 和 Inc-Cluster 通过衡量节点间的相似程度并进行聚类，使用了表示节点两两关系的二维矩阵进行中间结果保存，因此导致其在 k 增大时迭代中间结果超出系统虚存承载能力。在大规模的数据集上，由于采用了阈值算法控制输出结果，因而获得了较好的伸缩性。

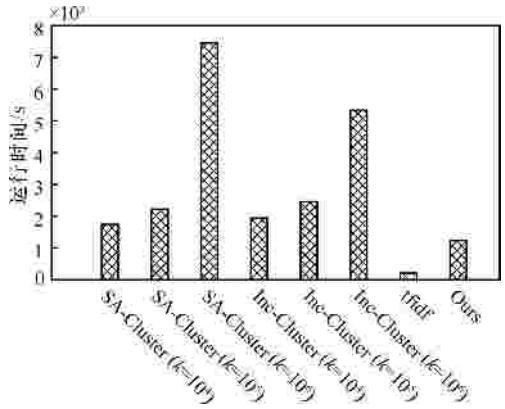


图 13　运行时间分析

7　结束语

社会网络是近年来的研究热点之一，本文给出了一种社会网络上特征簇的识别和提取方法，并在真实数据集上进行了验证。簇的特征在某种意义下，可以作为理解社区形成机理，挖掘特定人群特殊兴趣等方面的依据。研究特征簇的挖掘技术在商业、学术甚至军事上都有具有广泛的应用前景，例如在社交网络商业推广、知名学术团体挖掘、军用传感网情报挖掘等。

本文的工作是在静态的网络结构上进行的特征簇挖掘，目前，许多社会网络的结构是随时间演变的，演化网络上的时间特征也是一个非常值得研究的方向，例如本文的后续工作就可以在分析簇的演化规律，挖掘时间相关的特征簇等方面开展。

参考文献：

[1] NEWMAN M E J, WATTS D J, STROGATZ S H. Random graph models of social networks[A]. Proceedings of the National Academy of Sciences of the United States[C]. America, 2002.2566.

[2] ZHOU D, ORSHANSKIY S A, ZHA H Y, *et al*. Co-ranking authors and documents in a heterogeneous network[A]. Proceedings of the 2007 IEEE International Conference on Data Mining (ICDM'07)[C]. 2007. 739-744.

[3] PAGE L, BRIN S, MOTWANI R, *et al*. The Pagerank Citation Ranking: Bringing Order to the Web[R]. Stanford University, 1998.

[4] YAN X F, HAN J W. Gspan: graph-based substructure pattern mining[A]. Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)[C]. 2002. 721.

[5] ZHOU Y, CHENG H, YU J X. Graph clustering based on structural/attribute similarities[J]. PVLDB, 2009, 2(1):718-729.

[6] HANNEMAN R A, RIDDLE M. Introduction to Social Network Methods[M]. University of California, Riverside, 2005.

[7] DOURISBOURE Y, GERACI F, PELLEGRINI M. Extraction and classification of dense communities in the Web[A]. WWW[C]. 2007. 461-470.

[8] ZENG Z P, WANG J Y, ZHOU L Z, *et al*. Coherent closed quasi-clique discovery from large dense graph databases[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06[C]. New York, NY, USA, 2006. 797-802.

[9] KARP R M. Reducibility among combinatorial problems[A]. Complexity of Computer Computations[C]. New york, 1972.

[10] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)[C]. 2006. 611- 617.

[11] PALLA G, DERÉNYI I, FARKAS I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043):814-818.

[12] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69:26113.

[13] FLAKE G, LAWRENCE S, GILES C L. Efficient identification of web communities[A]. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)[C]. 2000. 150-160.

[14] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74(3): 036104.

[15] **沈华伟，程学旗，陈海强等. 基于信息瓶颈的社区发现**[J]. **计算机学报**, 2008, 31(4):677-686.
SHEN H W, CHENG X Q, CHEN H Q, *et al*. Information bottleneck based community detection in network[J]. Chinese Journal of Computers, 2008, 31(4):677-686.

[16] INOKUCHI A, WASHIO T, MOTODA H. An apriori-based algo-

rithm for mining frequent substructures from graph data[A]. Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)[C]. 2000. 13-23.

[17] KURAMOCHI M, KARYPIS G. Grew-a scalable frequent subgraph discovery algorithm[A]. ICDM[C]. 2004. 439-442.

[18] VANETIK N, GUDES E, SHIMONY S E. Computing frequent graph patterns from semistructured data[A]. ICDM[C]. 2002. 458-465.

[19] HUAN J, WANG W, PRINS J. Efficient mining of frequent subgraphs in the presence of isomorphism[A]. ICDM[C]. 2003. 549-552.

[20] HUAN J, WANG W, PRINS J, *et al*. Spin: mining maximal frequent subgraphs from graph databases[A]. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, KDD '04[C]. New York, NY, USA, 2004. 581-586.

[21] PORTER M F. An Algorithm for Suffix Stripping[M]. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1997. 313-316.

[22] ZHOU Y, CHENG H, YU J X. Clustering large attributed graphs: an efficient incremental approach[A]. ICDM[C]. 2010. 689-698.

[23] TIAN Y Y, HANKINS R A, PATEL J M. Efficient aggregation for graph summarization[A]. SIGMOD Conference[C]. 2008. 567-580.

作者简介：

韩毅（1982-），男，蒙古族，内蒙古奈曼旗人，博士，国防科技大学计算机学院助理研究员，主要研究方向为数据挖掘和信息获取。

方滨兴（1960-），男，江西万年人，中国工程院院士，北京邮电大学校长，国防科技大学特聘教授，主要研究方向为网络信息安全。

贾焰（1960 -），女，四川成都人，国防科技大学教授、博士生导师，国防科技大学网络与信息安全研究所副所长，主要研究方向为数据挖掘技术和网络信息安全。

周斌（1971-），男，江西南昌人，国防科技大学教授、硕士生导师，主要研究方向为中间件技术和网络信息安全技术。

韩伟红（1973-），女，吉林长春人，国防科技大学副教授、硕士生导师，主要研究方向为数据库技术和网络信息安全技术。

（上接第 37 页）

[15] TIAN H, SHEN H, MATSUZAWA T. Developing energy-efficient topologies and routing for wireless sensor networks[A]. Proc of International Conference on Network and Parallel Computing[C]. Beijing, China, 2005.

[16] GILBERT E N. Random graphs[J]. Ann Math Statist, 1959, 30(4): 1141-1144.

作者简介：

方效林（1984-），男，江西上饶人，哈尔滨工业大学博士生，主要研究方向为传感器网络协议及数据处理。

高宏（1966-），女，博士，黑龙江哈尔滨人，哈尔滨工业大学教授，主要研究方向为并行数据库、并行压缩数据仓库、数据流、传感器网络数据处理。

熊蜀光（1982-），男，重庆人，哈尔滨工业大学博士生，主要研究方向为传感器网络查询与数据处理。