社交网络的结构支撑理论

韩 毅[1),2)] 许 进[1)] 方滨兴[2),3)] 周 斌[2)] 贾 焰[2)]

[1)](北京大学计算机科学与技术系 北京 100871)

[2)](国防科学技术大学计算机学院 长沙 410073)

[3)](北京邮电大学计算机学院 北京 100876)

摘 要 社交网络分析是近年来的研究热点之一,常见的分析方法包括度分布分析、个体排名、社区发现、模式发现等.文中,作者认为一个人的社会地位与其所在的网络结构具有紧密的联系,而这种网络结构对成员社会地位的影响程度是可以被表示和量化的.文中通过分析社交网络的链接结构,将社交网络中个体与个体间的依赖关系从一般社会关系中抽取出来,提出了一种基于依赖模型的支持力衡量方法,并基于此给出了一种高效的计算最具支持力的节点计算方法.此外,基于上述模型,设计了一种基于依赖关系的支撑结构模型及其计算方法,用于刻画社交网络中特定节点的影响力来源.作者在大规模的真实数据环境下对模型和算法的正确性、效率和伸缩性进行了验证.

关键词 社交网络;依赖度;支持力;社区发现

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.00905

Structural Supportiveness Theory on Social Networks

HAN Yi[1),2)] XU Jin[1)] FANG Bin-Xing[2),3)] ZHOU Bin[2)] JIA Yan[2)]

[1)](*Department of Computer Science, Peking University, Beijing* 100871)

[2)](*College of Computer, National University of Defense Technology, Changsha* 410073)

[3)](*School of Computer, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract Social network analysis has been studied extensively from variable angles such as degree distribution analysis, social entity ranking, community extraction, and pattern discovery, etc. In this paper, we consider a person's social status is highly related with the network structure which he/she locates in, and such impact from network structure to entities' social status can be modeled and measured. We extract the dependencies from ordinary relations by analyzing the link structure. We also propose a "supportiveness" model based on dependency model. We exploit a supportiveness-based entity ranking scheme, and efficient algorithms are developed to compute the top-n most supportive entities. Moreover, we extend the supportiveness analysis to community extraction, and develop feasible solutions to identify the most supportive groups of entities. The empirical study conducted on a large real data set indicates that the supportiveness measures are interesting and meaningful, and our methods are effective and efficient in practice.

Keywords social network; dependency; supportiveness; community extraction

收稿日期:2013-06-21;最终修改稿收到日期:2014-01-24.本课题得到国家"九七三"重点基础研究发展规划项目基金(2013CB329601,2013CB329602)、国家自然科学基金(61372191,91124002,61133010,61033003)、博士后科学基金(2013T60037,2012M520114)资助.
韩 毅,男,1982 年生,博士,高级工程师,中国计算机学会(CCF)会员,主要研究方向为信息网络分析、数据挖掘、信息安全.E-mail:yihan@pku.edu.cn.**许 进**,男,1959 年生,博士,教授,主要研究领域为图论、DNA 计算机和信息安全等.**方滨兴**,男,1960 年生,博士,教授,中国工程院院士,主要研究领域为信息网络分析、信息安全和网络安全等.**周 斌**,男,1971 年生,博士,教授,主要研究领域为信息网络分析和数据挖掘等.**贾 焰**,女,1960 年生,博士,教授,主要研究领域为信息网络分析、数据挖掘、信息安全.

1 引 言

近年来,在线社交的快速发展对个人工作生活、国家经济发展、社会稳定和国家安全都会产生新的影响作用.特别是以微博为代表的在线社交网络在新闻传播、促进信息公开、网民互帮互助、微博品牌营销、传播信息和知识等方面都表现出了非常积极的作用.然而,在社交网络全面进入人们生活的同时,也可以发现微博等社交网络上虚假谣言信息盛行、网络推手和欺诈活动猖獗,甚至针对社交网络的蓄意的煽动破坏等现象也给正常的互联网社会活动带来了有害的影响.

通过不断爆发的舆情事件可以看出,一方面,对于快速传播的网络信息内容,社交网络的小世界现象和级联式传播方式具有很强的催化放大作用,极大地加速着信息的传播和演化;另一方面,社交网络用户通过在阅读、交友和游戏过程中传播信息,也在不断地淘汰低质信息源,生成新的社会关系和信息通道,这种对信息源的再选择行为也导致了网络结构的变化.

社交网络用户利用网络拓扑结构进行信息传播和影响力放大,本文中我们将解读这种网络个体利用网络拓扑结构相互作用这一过程."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?"、"一个链接密集的结构簇是一群朋友、亲戚还是对于特定话题的兴趣小组?"、"如何识别既有的网络推手组织?".概括而言,对上述问题的研究可以归纳转化为以下研究问题:即(1)一个特定网络结构是否和其承载的信息相关,这种相关性是否可以量化计算?(2)给定一类信息内容(特征),如何识别出与其相关的特定网络结构(推手),例如带有一定特征的节点簇?(3)在线社交网络的网络结构如何影响内容的演化?

本文中的社交网络的结构支撑理论主要研究社交网络中的节点利用网络拓扑结构进行相互作用和影响这一过程.社交网络中,成员往往在信息交换中体现出一种相互依存的特点.举例来说,微博网络中营销账号的影响力多来自于网络水军构建相应的特定传播结构,而有些知名ID走红的原因则来源其个人固有属性,例如明星效应、自身发表内容的吸引力等.相比起来,前者对社交网络本身表现出一种强烈依赖的特征,即如果将其支持者与其关系断开,其受关注程度必然会有大幅下降;而对于后者,网络结构对其的影响力相对较小.

这一现象是由社会成员在网络中体现出的社会性造成的.当网络结构发生变化时,网络个体的自身影响力也发生改变,则这种改变可以被理解成个体对其网络结构的依赖.简单来说,在一个社交网络中,如果将节点 v 断开,会导致另外一个节点 u 的社会影响力发生变化,则我们认为 u 在某种程度上依赖于 v.将 u 的社会地位变化的程度量化表示,就是 u 对 v 的依赖度.反过来看,对于两个节点 u 和 v,u 对 v 的依赖可以理解为 v 对 u 的支持,而一个节点对于网络上其他节点的支持程度总和,就是这个节点对网络的支持力.

研究社交网络的结构支撑理论在信息检索、学术引用/合作查询、在线购物、在线评价等应用中,都具有广泛的研究前景.例如可以通过依赖度和支持力的研究作为发现关键成员的依据,与传统的节点影响力函数不同,个体的支持力更关注于个体对网络的贡献程度,更加体现了一种"给力"的思想.

在过去的社交网络研究中,类似的研究还不多见,本文在分析了相关研究的基础上,做出了以下贡献:

首先,在分析了社交网络中影响力度量函数原理的基础上,给出了个体间依赖度的形式化定义,将一般社交网络结构转化为依赖度网络.

其次,在依赖度网络模型的基础上,给出了基于依赖度的个体间距离定义方法,并结合最近邻和 k 近邻方法,给出了一种类似于"投票"的个体支持力指标,并设计了高效的计算方法.

再次,我们将依赖模型扩展到社交网络的社区发现技术上来.我们通过依赖度模型,发现对特定成员影响力构成起到重要作用的网络群体,即支撑结构.

最后,我们在真实的数据集上对本文提出的算法进行了验证,实验结果说明我们的模型是正确的,结果是有意义的,且算法是高效的.

2 相关工作

目前专门针对社交网络中节点间依赖关系的挖掘和分析的研究工作还不多见,本文在给出了节点间的依赖模型的基础上,重点研究了基于依赖模型的个体支持力分析和社区发现.

个体支持力指的是社交网络中的节点对他人提供支持的力度的一种衡量,本文中我们设计的支持力模型是一种基于社交网络结构分析的重要性指标.通过分析网络结构对个体进行重要性指标衡量,

近年来在研究上获得了广泛的关注，其中著名的例子之一要数上世纪末提出的随机游走模型和PageRank模型[1]. 主要思想是将用户浏览网页的行为模型化为在网页链接结构中根据链接方向进行随机前进，并具有一定的概率随机跳转到其他页面. 由于网络的链接疏密程度和复杂网络中呈现的小世界模型[2]，每个页面在随机游走模型下获得访问的概率也不尽相同，这种概率就是PageRank的思想. PageRank将链接关系的结构提炼，把这种结构带来的信息传递效应转化为节点的重要性指标，与之类似的方法还有康奈尔大学的Kleinberg等人[3]提出的HITS模型等. PageRank根据其所在网络特点和分析目标的不同，也产生了一些变种，例如，判断节点间距离的RWRS[4]以及结合话题主题的随机游走[5]等. 本文在类似的重要性定义的基础上，分析链接结构对节点重要性的影响，即个体对网络结构的依赖度. 本文还提出了一种支持力度量的定义，与上述重要性度量不同的是，支持力度量则更侧重于个体对于整体的贡献.

在目前的社区发现技术中，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[6-7]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，Pattillo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于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等人[14]提出了模块性的概念，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，中国科学院计算技术研究所的沈华伟等人[15]提出了使用信息瓶颈模型来发现社区，通过寻找网络的最优压缩表示来发现网络的社区结构. 本文中我们的目标之一是要挖掘一种支撑结构，即社区的成员间应互相依存，密不可分，上述提到的方法与本文的侧重不同，都不能直接应用于本文的场景.

3　依赖性和支持力计算模型

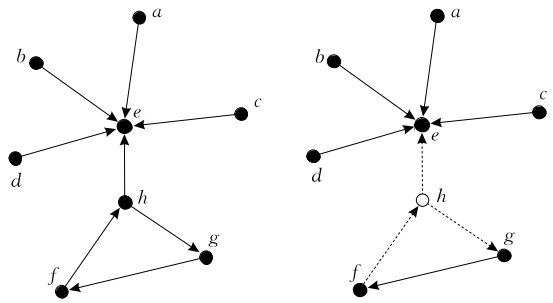
现实生活中，许多社交网络，例如在线交友网站、学术合作网络、通信网络、生物蛋白质相互作用网络，都可以被模型化为图. 本文中，社交网络由二元组$G=\langle V,E\rangle$表示，其中$v\in V$表示个体和个体集，个体间的关系由边$e=\langle u,v\rangle\in E$表示，$E$代表边(链接)集.

为了度量社交网络上的节点在社交网络中的地位，我们引入节点的影响力度量函数：$\lambda:V\rightarrow R$，它将节点集映射为一组实数，$\lambda_G(v)$表示节点v在图G上的影响力度量函数取值.

3.1　节点的影响力度量函数

本文中提到的影响力度量函数$\lambda_G(v)$是一个抽象函数，任何社交网络上相关的节点重要性打分函数，例如PageRank、TwitterRank[16]等都可以作为λ的实现.

例1.　节点间的依赖关系. 图1表示一个由8个节点组成的社交网络，表1第1行表示对应节点的PageRank计算取值. 从表1中可以看出，对于节点e，$PR_G(e)$表示其在图G中的PageRank取值. $G-h$表示当在图G中断开节点h后的图(图1(b)). 明显的，当分别断开f，g或h时，$PR_G(e)$发生了超过50%的变化，然而去掉节点a，b，c或d时，$PR_G(e)$却没有明显改变.



(a) 原图　　(b) 断开节点h

图1　一个8节点图的例子

表1　图1所示图结构的影响力度量函数取值

	a	b	c	d	e	f	g	h
PR_G	2.44	2.44	2.44	2.44	27.39	20.79	16.56	25.48
PR_{G-a}		2.45	2.45	2.45	25.28	21.47	17.10	26.33
PR_{G-b}	2.45		2.45	2.45	25.28	21.47	17.10	26.33
PR_{G-c}	2.45	2.45		2.45	25.28	21.47	17.10	26.33
PR_{G-d}	2.45	2.45	2.45		25.28	21.47	17.10	26.33
PR_{G-e}	2.05	2.05	2.05	2.05		29.91	29.91	29.91
PR_{G-a}	5.11	5.11	5.11	5.11	58.33		11.02	5.11
PR_{G-a}	4.28	4.28	4.28	4.28	61.78	4.28		12.56
PR_{G-a}	5.11	5.11	5.11	5.11	52.42	16.94	5.11	

这就说明,当网络的结构发生变化时,即使发生变化的部分不与给定节点 v 直接相连,也可能会对 v 的重要性产生影响.这种结构对个体重要性的影响就是节点对结构的依赖.

由于节点的重要性取值通过对链接结构的分析计算得到,因此链接结构的改变必然会造成节点影响力的变化.

定义 1. 依赖度.对于一个图 $G=\langle V,E\rangle$,个体 $u,v\in V$ 且 $u\neq v$,影响力度量函数 λ_G,我们将 u 对 v 的依赖度函数 $\varpi(v\to u)$ 定义为

$$\varpi(v\to u)=\frac{|\lambda_G(u)-\lambda_{G-v}(u)|}{\lambda_G(u)},$$

其中,$G-v$ 表示在图 G 中去掉节点 v 之后的导出子图,$\lambda_G(v)$ 表示个体 v 在网络 G 下的重要性取值.$\varpi(v\to u)$ 表示了图 G 去掉 v 后 u 重要性产生的变化在原有重要性得分中的比重;反过来看,$\varpi(v\to u)$ 也可以理解为 v 对 u 的支撑程度.

直观的理解,对于一个个体 v,如果去掉网络中的一个子部分会使 v 的重要性受到较大的变化,则可以认为 v 对这个子部分具有较大程度的依赖,这种依赖程度可以被理解为衡量节点间亲疏关系的一种标准.依赖越大,表示节点间关系越紧密;依赖越小,表示节点间关系越疏远.值得注意的是,根据 λ_G 的具体实现函数的不同,ϖ 的取值范围也会有不同,对于某些基于访问概率的函数,例如 PageRank,节点总数的改变会对节点影响力的平均值产生影响.

值得特别注意的是,我们在计算 $\lambda_{G-u}(v)$ 时,并不真的将 u 从 G 中删除,而是断开 u 与其他节点的连接,将其置于孤立位置,从而确保节点总数不变.很容易推断出,网络中本身处于孤立位置的节点,对任何其他的节点的依赖度都为 0.

依赖度可以作为一种节点间的距离衡量标准,即两个节点相互依赖程度越高,其距离就越短.由于网络链接空间本身就可以被视为一个立体的、扭曲的空间,因此这种紧密度距离并不在可度量空间上,即其任意 3 个两两相邻的节点间的紧密度距离并不符合三角定律.

计算依赖度最直观的方式就是对于每个 $v\in V$ 首先计算 $\lambda_G(v)$,之后通过断开节点 u 的链接从而获得 G_u,并计算 $\lambda_{G-u}(v)$ 并进一步计算 $\varpi(u\to v)$.然而,在大规模的网络下,这样的计算方法将耗费大量的计算资源.实际上,精确计算 $\varpi(u\to v)$ 也不必要也不现实.实验证明,局部链接的改变并不会影响全局,即 PageRank 的衰减效应[17].Langville 等人[17] 给出了一种通过分析单个链接来估算 PageRank 的方式.

3.2 个体的支持力模型

依赖度可以被理解为衡量节点间亲疏关系的一种标准,依赖越大,表示节点间关系越紧密;依赖越小,表示节点间关系越疏远.

通过依赖度,我们可以将依赖度转化为节点间的相对距离.我们将本章中图的表示方式转化为二元组 $G=\langle V,E_\varpi\rangle$ 表示,其中 $\varpi:V\times V\to R$ 是节点间的依赖度函数,即对于任意一个节点对 $\langle u,v\rangle$,其中 $\varpi(v\to u)$ 表示 u 对 v 的依赖度取值,$E_\varphi=\{\langle u,v\rangle\mid \varpi(v\to u)\neq 0\}$ 表示相互具有依赖关系的节点对.

对于两个节点 u 和 v,u 对 v 的依赖可以理解为 v 对 u 的支持,因此,一个节点被其他节点依赖的程度,可以被理解为这个节点的“给力”程度,即节点的支持力.

为了量化计算社交网络中个体间的联系,我们设计了一种基于 k 近邻关系的支持力指标.

定义 2. 支持力.在依赖关系图 $G=\langle V,E_\varpi\rangle$ 中,对于 $v\in V$,v 的最近邻集合,表示为 $NN(v)$,有

$$NN(v)=\{u\in V\mid \nexists u'\in V:\varpi(u'\to v)>\varpi(u\to v)\}.$$

v 的反向最近邻,表示为 $RNN(v)$,有

$$RNN(v)=\{u\mid v\in NN(u)\}.$$

在图 $G=\langle V,E_\varphi\rangle$ 中,$u\in V$ 的支持力,表示为 $supp(u)$,有

$$supp(u)=\sum_{v\in RNN(u)}\frac{1}{|NN(v)|}\tag{1}$$

k 近邻是一种使用非常广泛的关系;例如,在在线交友的网站中,网站一般会根据用户的要求,同时返回最合乎用户理想的多个候选人,供用户挑选.最近邻 $NN(u)$ 表示 u 最依赖的节点在 $NN(u)$ 之外,不存在 u 依赖更多的节点.$RNN(u)$ 表示依赖 u 最多的个体集.如果一个节点 u 的 $RNN(u)$ 集合元素数量越大,则说明其受到越多的个体依赖.值得注意的是,通常情况下 $|NN(v)|=1$,表示每个节点只有一个最近邻,所以有 $supp(u)=|RNN(u)|$.如果 v 同时视多个节点为最近邻,即 v 对 $x,y,\cdots$ 等 n 个节点依赖度相同且最大,才会有 $|NN(v)|\geq 2$,公式 $\sum_{v\in RNN(u)}\frac{1}{|NN(v)|}$ 表示视 u 为最近邻的节点数.当 $|NN(v)|\geq 2$ 时,为保证公平,v 对 $x,y,\cdots\in NN(v)$ 的贡献值为 $\frac{1}{|NN(v)|}$.

与 $NN(u)$ 定义类似,$kNN(u)$ 代表 u 的 k 近邻集合.对于自然数 $k\geq 1$,$v\in kNN(u)$ 表示无法找到 k 个 $v'\in V$ 使 v' 满足 $\varpi(u'\to v)>\varpi(u\to v)$.其意义代表 $kNN(u)$ 是节点 u 在网络 G 中最近的 k 个节点.值

得注意的是，如果$u\in kNN(v)$，则$u\in(k+1)NN(v)$；反之不然.

定义 3.　k-支持力. 在图$G=\langle V,E_\varpi\rangle$中，节点$u\in V$的$k$支持力定义为

$$supp_k(u)=\sum_{v\in RkNN(u)}\frac{k}{|kNN(v)|} \tag{2}$$

计算$supp(u)$的过程可以想象成为一个投票的过程，每个节点一张票，且都投给其最依赖的节点. 当一个节点有大于一个最依赖的节点，则为了公平起见，其选票将被分割给所有他最依赖的人. 类似的，在k支持力的计算中，每个个体可以投票k次，如果存在并列第k的情况，他需要将其k张票平均分给所有符合kNN中的个体.

3.3　支持力计算

给定一个图$G=\langle V,E_\varpi\rangle$和一个参数$k$，我们计算使支持力函数$supp_k(u)$取值最大的$n$个节点.

直观的方法是，对于每个节点$u\in V$，首先计算$kNN(u)$；给每个$v\in kNN(u)$投票$\frac{1}{|NN(v)|}$次；最后统计每个节点的得票数量，并输出 top-n. 算法 1 描述了这一过程的伪代码.

算法 1.　计算支持力最大的n个节点.

输入：依赖关系图$G=\langle V,E_\varpi\rangle$，参数$n$和$k$；

输出：计算支持力最大的n个节点；

1. FOR $v\in V$
2. 　　在依赖关系图中扫描v的邻居并计算$kNN(v)$；
3. 　　对于每个$u\in kNN(v)$，对u记$\frac{k}{|kNN(v)|}$票；
4. ENDFOR
5. FOR $v\in V$
6. 　　计算v收到的总票数；
7. ENDFOR
8. 输出得票最高的n个节点；

4　支撑结构的性质和计算方法

个体间的依赖度函数用于度量个体与个体单向关系. 直观理解，对于一个节点v，如果在社交网络中存在一组节点S，使v紧密依赖于S，则这组节点可以看做是这个节的支撑结构. 也就是说，S对v的影响力起着非常重要的作用.

定义 4.　社区支撑度. 在图$G=\langle V,E\rangle$中，一组节点$S\subset V$和一个节点$v\notin S$，以及节点影响力度量函数λ，S对v的支撑度，表示为

$$\varpi(S\to v)=\frac{|\lambda_G(v)-\lambda_{G-S}(v)|}{\lambda_G(v)} \tag{3}$$

直观理解，S的对v的支撑度表示S对v的影响力的贡献程度，反过来看，S的对v的支撑度$\varpi(S\to v)$也表示v对S的依赖度. $\varpi(S\to v)$与S对v的贡献成正比.

4.1　支撑结构的性质

支撑结构的性质与节点影响力度量函数λ紧密相关，本节我们以 PageRank 为例，分析支撑结构的性质.

PageRank 是一种度量有向网络上节点影响力的方法. PageRank 利用随机游走的思想，将每个节点在随机游走过程中受访的概率作为度量节点影响力的指标.

$$\lambda_{PR}(u)=\frac{1-\varepsilon}{|V|}+\varepsilon\sum_{v\in M(u)}\frac{\lambda_{PR}(v)}{d(v)} \tag{4}$$

PageRank 利用一种概率转移的思想：节点v将自己的 PageRank 取值均分，转移给其指向的节点；一个节点u的 PageRank 取值$\lambda(u)$取决于其受到的转移量之和. 其中，ε表示阻尼因子，表示一定概率的随机跳转，并用于保证迭代过程收敛；$M(u)$表示指向u的节点集合，$d(v)$表示节点v的出度，由于$v\in M(u)$，因此一定有$d(v)>0$.

Brinkmeier[18] 通过对 PageRank 公式(4)进行变形展开得到如下性质：

对于G中的路径$P=v_0\to\cdots\to v_n\to u$，路径$P$对$u$的贡献表示为

$$\varpi(P\to u)=\frac{1}{|V|}\varepsilon^{n+1}(1-\varepsilon)\prod_{i=0}^{n}\frac{1}{d(v_i)} \tag{5}$$

节点u在图G中的 PageRank 则可以表示为

$$\lambda_{PR}(u)=\frac{1-\varepsilon}{|V|}+\sum_{v\in C(u)}\Big(\sum_{P\in D(v,u)}\varpi(P\to u)\Big) \tag{6}$$

其中，$C(u)$表示有路径通向节点u的节点集合. $D(v,u)$表示v和u之间的有向路径集合.

定理 1.　支撑度的单调性. 对于图$G=\langle V,E\rangle$，G上的 PageRank 度量函数$\lambda_{PR}:V\to R^+$，如果有两个节点集$S\subset S'\subset V$且满足$v\notin S,S'$，社区支撑度函数ϖ满足$\varpi(S'\to v)\geq\varpi(S\to v)\geq 0$.

证明.　为了证明$\varpi(S'\to v)\geq\varpi(S\to v)\geq 0$，我们首先分析节点$v_0\in S'-S$的情况.

如果$v_0\notin C(v)$，则说明v_0与节点v无连接关系，则根据式(6)可知，$\varpi(v_0\to v)=0$；

如果$v_0\in C(v)$，则必然存在路径$v_0\to\cdots\to v_n\to v$，且$v_0\notin S$，则必然可以推出$\varpi(v_0\to v)>0$.

根据下文的定理 2，$\varpi(v_0\to v)+\varpi(S\to v)\geq\varpi(S\cup\{v_0\}\to v)\geq\varpi(S\to v)$，从而可以推出，$\varpi(S'\to v)\geq\varpi(S\to v)\geq 0$.　　证毕.

定理 2.　支撑度的估计. 对于图$G=\langle V,E\rangle$，G上的 PageRank 度量函数$\lambda_{PR}:V\to R^+$，对于任意两

个集合 $S_1, S_2 \subset V$，有

$$\varpi(S_1 \to v) + \varpi(S_2 \to v) \geq \varpi(S_1 \bigcup S_2 \to v) \quad (7)$$

证明. 式(6)表明，一个节点 u 的 λ_{PR} 取值仅和与 u 有直接路径的节点有关. 对于节点 $v_1 \in S_1$，$v_2 \in S_2$，且 $v_1 \neq v_2$，根据式(6)，显然有 $\varpi(v_1 \to v) + \varpi(v_2 \to v) \geq \varpi(\{v_1, v_2\} \to v)$，因此显然可推导出式(7). 证毕.

值得注意的是，上述两个定理对 PageRank 的各类变种也都适用.

4.2 支撑结构的计算方法

给定一个节点，如何高效计算其支撑结构？以 PageRank 为例，根据式(6)，一个连通子图中任何有路径通向节点 v 的其他节点都会为其提供支撑. 一个直观的定义可以对于一个节点 v，是设定一个门槛值 θ，计算一个最小的节点集 S，使 $\varpi(S \to v) > \theta$.

然而不难看出，上述问题可以对应为经典的背包问题，也就是说，θ 支撑结构的计算是一个 NP 难问题.

为了简化这一过程，也为了使支撑结构能够验证社交网络中信息传播过程，在实践中我们采用了一种近似的方法(算法 2).

算法 2. 计算 θ 支撑结构.

输入：关系图 $G=\langle V, E\rangle$，参数 θ，节点 $v \in V$

输出：支撑结构 S，满足 $\varpi(S \to v) > \theta$；

1. $S=\{v\}$；
2. WHILE true
3. IF $\varpi(S \to v) > \theta$；
4. BREAK；
5. ENDIF
6. 让 $M(S)$ 表示指向 S 的节点集合；
7. $S = S + \arg\max\{\varpi(u \to v) \mid u \in M(S)\}$；
8. ENDWHILE

首先从 v 开始，将直接指向 v 的节点中对 v 支持最大的节点加入 S，并继续在指向 S 的节点中搜索，直到结果满足预设条件.

值得注意的是，算法 2 中找到的结果并不满足使 S 最小这一条件. 然而由于每轮都是在相邻的节点中搜索，这样找到 S 时必然是一个连通子图. 在微博的消息传递网络中，这样的支撑结构能够体现转发和评论行为在信息传播过程中起到的放大作用的程度.

5 实验分析

我们在真实的微博数据集上验证了本文的模型和算法的有效性，并在一个更大规模的数据集上测试了算法的效率.

算法使用 Java 实现，具体软件和硬件环境如表 2 所示.

表 2 实验环境

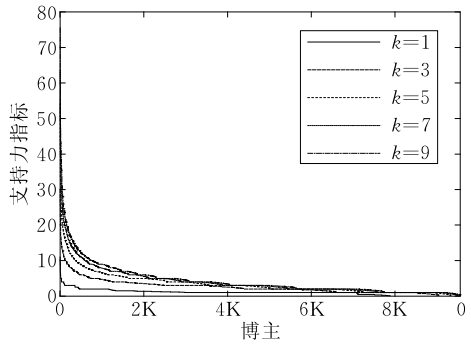
CPU	Intel Core i7 3 GHz
内存	16 GB DDR3
操作系统	CentOS 内核 2.6.18
运行环境	Java Runtime Environment 1.6

5.1 有效性验证

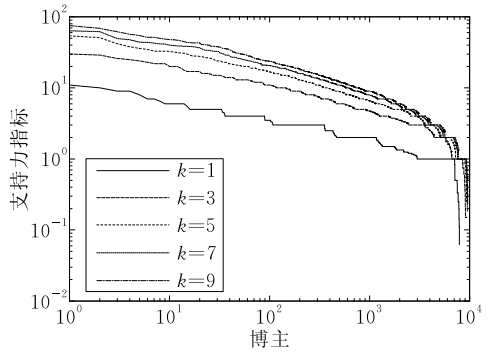
我们在 2013 年 6 月抓取的新浪微博的相互转发关系网络数据上验证了本文提出的模型和方法的有效性.

数据包含 10 427 个博主和其 24 313 次转发关系. 在该数据集中，每个节点代表新浪微博的一个博主 ID，如果一个博主转发了另一博主的一条微博，那么他们之间就会存在一条有向边. 我们称该数据集为 SN 数据. 我们考虑微博中关注关系网是一种基于“订阅-分发”模式的内容信息通路，而微博中的评论行为则表示用户对特定内容的反馈，评论和关注并不会直接造成信息的传播和信息影响力的放大；与关注和评论行为不同，转发行为可以造成信息的二次传播，更能体现影响力的放大和“支撑”的效果，因此我们选用了转发网络作为实验验证.

图 2 表示了所有节点在不同 k 值约束下的支持



(a) 标准刻度

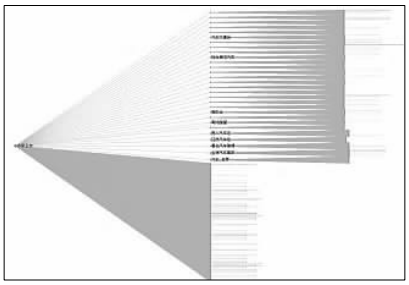


(b) 对数刻度

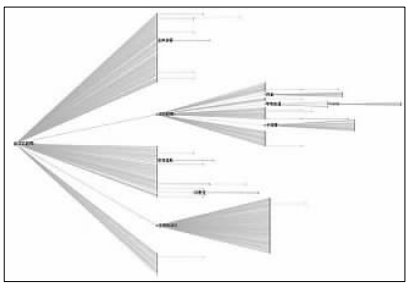
图 2 节点的支持力分布

力分布. 支持力的定义是一种基于依赖度的反向 k 近邻的定义，从图中可以看出，虽然在计算支持力时并没有要求节点直接连接，节点支持力的分布表现出了一种幂律分布的特征.

为了验证支撑结构的有效性，我们选取了特定博主和博文进行支撑结构的分析. 图 3 是 UID 分别为 1276821910 和 2145494291 的两个用户的支撑结构(2013 年 6 月). 可以看出，图 3(a)所示用户转发结构非常简单，且层级很少，形状规则；而图 3(b)所示用户的结构关系复杂，层级也较多. 可以推断，图 3(a)为利用水军构建的人工转发结构.



(a) UID: 1276821910



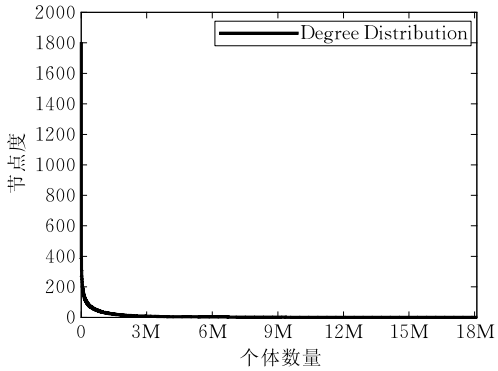
(b) UID: 2145494291

图 3　新浪微博中两个真实节点的支撑结构关系图

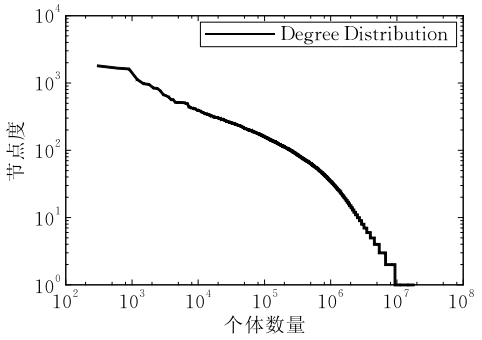
5.2　算法效率分析

受采集手段和站点访问的限制，本文使用的新浪数据集的规模较小，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的效率的影响，我们于 2010 年 4 月在境外 Twitter 网站上通过其官方 API 抓取了公开 ID 的一个连接构件(connected component)用于测试性能和伸缩性. 在这个数据集中，每个注册 ID 表示为一个节点，ID 间的 WFW(Who Follows Whom，Twitter 中的关注)关系构成有向边集. 共计包含 18012823 个注册 ID 和其对应的 33237699 个 WFW 关系，本文中称该数据集为 TW 数据集. 通过图 4 可以看出，数据集在节点度分布上都呈长尾形式.

我们利用 TW 数据分析前文所述算法的效率和伸缩性. 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整个网络进行划分，递增式地爬取了 5 个不同规模的子部分，用于测试算法在不同节点/链接规模下的伸缩性. 图 5(a)展示了每个部分节点数和边数的分布情况，图 5(b)展示了 θ 支

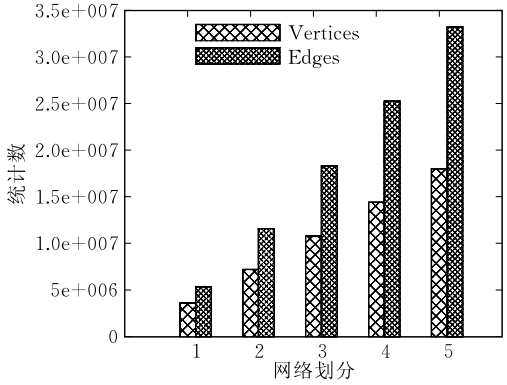


(a) 标准刻度

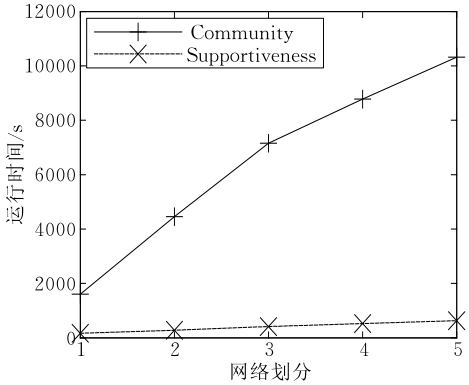


(b) 对数刻度

图 4　节点度分布：Twitter 数据集



(a)



(b)

图 5　节点数与运行时间

撑结构($\theta=0.9$)的平均计算时间和支持力排名计算的运行时间.

由于节点规模较大,内存中不能完全存储社交网络的邻接表.由于支撑结构计算需要大量多次访存邻接表,因此支撑结构计算的平均时空开销明显高于支持力排名计算.可以看出,随着节点数量的增加,由于支持力和支撑社区计算都是在相邻节点中进行搜索,总运行时间都接近于线性时间,算法的伸缩性较好.

5.3 相关算法分析

与 PageRank、TwitterRank 等算法类似,本文的支持力度量本质上是一种节点的影响力度量方法;不同的是,支持力度量在原理上考虑的是节点的"被依赖能力",是基于 PageRank、TwitterRank 等影响力度量算法的改进.通过算法 1 可以看出,支持力算法的效率的关键在于计算影响力函数 $\lambda_G(u)$,以及基于此进一步计算依赖性 $\frac{|\lambda_G(u)-\lambda_{G-v}(u)|}{\lambda_G(u)}$ 和最近邻.

本节中,我们将算法 1 中的过程分为两部分计时,计算影响力度量函数 $\lambda_G(u)$ 计算部分占全部算法的比重.我们在 TW 数据集上分别测试了 3 种 $\lambda_G(u)$ 的实现占用的运算时间,这 3 种分别是社交网络结构属性度量方法节点介数(Node Betweenness)、基于随机游走的 PageRank 和考虑话题因素的 TwitterRank.

通过表 3 可以看出,影响力函数 $\lambda_G(u)$ 的计算占用了时间开销的绝大部分时间,除去 $\lambda_G(u)$ 的计算时间,3 种算法下支持力的计算时间基本相当,均在 50 分钟到 1 小时之间.结合算法 1 综合分析,支持力计算依赖网络中进行最近邻分析,当依赖网络已经生成完毕后,计算开销并不大.

表 3 支持力计算开销分析

	运行总时间/s	$\lambda_G(u)$运算占比/%
Node Betweenness	78328	93
PageRank	10346	67
TwitterRank	45873	91

我们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 TW 数据集上的运行时间进行了比较.SA-Cluster[19] 是一种综合考虑节点属性和结构特性网络节点聚类方法,通过自适应地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[20] 是 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,因此上述算法效率都对 k 的设置较为敏感.我们分别测试了 $k=10^4,10^5,10^6$ 的情况.

通过图 6 所示的结果可以看出,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,并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,迭代中间结果明显超出系统虚存承载能力.在大规模的数据集上,由于我们预计算了影响力度量函数 $\lambda_G(u)$,并采用了阈值算法控制输出结果,获得了较好的效果.

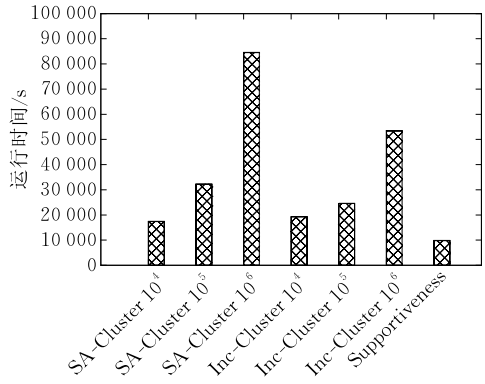


图 6 支撑社区运行时间分析

6 总结与展望

本文在分析了社交网络中节点影响力度量函数及其相互影响的基础上,给出了社交网络上依赖模型、支持力和支撑社区的定义、性质和计算方法,实验结果证明了本文的模型和方法是正确且高效的.社交网络上的结构支撑理论分析具有广泛的应用前景,例如,在在线社交网站上,通过分析节点在结构中的支持力,可以综合分析其承载的舆情信息的传播能力,从而可以进行商业推广,或进行舆论导向等研究.在 Web 结构上,对于一个给定节点,必定会有一组节点共同组成其支持社团,通过分析支持社团的互联结构,可以有效地识别人为构造的网络链接垃圾结构等.

在线社交网络是一个异质复杂网络,在社交网络的结构支撑理论方面,下一步可以将本文的工作进一步深化,研究社交网络中内容和结构的互依赖和支持关系,必将具有良好的研究前景.此外,如何利用网络大数据管理系统对本文提出的算法进行优化,也是一个下一步值得研究的问题.

参 考 文 献

[1] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Stanford University: Technical Report SIDL-WP-1999-0120, 1999

[2] Kleinberg J M. Navigation in a small world. Nature, 2000, 406(6798): 845-845

[3] Kleinberg J M, Kumar R, Raghavan P, et al. The Web as a graph: Measurements, models, and methods//Proceedings of the 5th Annual International Conference on Computing and Combinatorics. Tokyo, Japan, 1999: 1-17

[4] Fujiwara Y, Nakatsuji M, Onizuka M, Kitsuregawa M. Fast and exact top-*k* search for random walk with restart//Proceedings of the 38th Very Large Data Base. Istanbul, Turkey, 2012: 442-453

[5] Zhou Guangyou, Liu Kang, Zhao Jun. Topical authority identification in community question answering//Proceedings of the Chinese Conference on Pattern Recognition. Beijing, China, 2012: 622-629

[6] Hillery M, Reitzner D, Bužek V. Searching via walking: How to find a marked clique of a complete graph using quantum walks. Physical Review A, 2010, 81(6): 062324

[7] Jiang Daxin, Pei Jian. Mining frequent cross-graph quasi-cliques. ACM Transactions on Knowledge Discovery from Data(TKDD), 2009, 2(4): 16

[8] Pattillo J, Veremyev A, Butenko S, Boginski V. On the maximum quasi-clique problem. Discrete Applied Mathematics, 2013, 161(1-2): 244-257

[9] Richard M K. Reducibility among combinatorial problems//Junger M, et al. eds. 50 Years of Integer Programming 1958-2008. Berlin Heidelberg, Springer, 2010: 219-241

[10] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks//Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 337-357

[11] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818

[12] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 26113

[13] Flake G W, Lawrence S, Giles C L. Efficient identification of Web communities//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 2000: 150-160

[14] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006, 74(3): 36104

[15] Shen Hua-Wei, Cheng Xue-Qi, Chen Hai-Qiang, Liu Yue. Information bottleneck based community detection in network. Chinese Journal of Computers, 2008, 31(4): 677-685 (in Chinese)
(沈华伟，程学旗，陈海强，刘悦. 基于信息瓶颈的社区发现. 计算机学报，2008，31(4)：677-685)

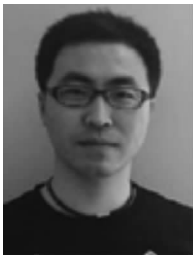
[16] Weng Jianshu, Lim Ee-Peng, Jiang Jing, He Qi. TwitterRank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 261-270

[17] Langville A N, Meyer C D. Survey: Deeper inside PageRank. Internet Mathematics, 2004, 1(3): 1-33

[18] Brinkmeier M. PageRank revisited. ACM Transactions on Internet Technology(TOIT), 2006, 6(3): 282-301

[19] Zhou Yang, Cheng Hong, Yu Jeffrey. Graph clustering based on structural/attribute similarities//Proceedings of the 35th Very Large Data Base. Lyon, France, 2009, 2(1): 718-729

[20] Zhou Yang, Cheng Hong, Yu Jeffrey. Clustering large attributed graphs: An efficient incremental approach//Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia, 2010: 689-698



HAN Yi, born in 1982, Ph. D., senior engineer. His research interests include information network analysis, data mining, information security.

XU Jin, born in 1959, Ph. D., professor. His research interests include graph theory, DNA computing, and information security.

FANG Bin-Xing, born in 1960, Ph. D., professor. His research interests include information network analysis, information security, and network security.

ZHOU Bin, born in 1971, Ph. D., professor. His research interests include information network analysis and data mining.

JIA Yan, born in 1960, Ph. D., professor. Her research interests include information network analysis, data mining, information security.

Background

A social network is a social structure made up of individuals which are tied by social links. In recent years, with the rapid development of information technology, online social networking services and micro blogging service received a lot of attentions. Social networks provide people a comprehensive communication platform of interaction, knowledge sharing, information dissemination, and so on. It also brought a significant impact on people's daily life and behaviors.

Social network analysis has been studied extensively from variable angles such as degree distribution analysis, social entity ranking, community extraction, and pattern discovery, etc. The proposed work is highly related with community detection problem on social networks. According to the definition of community structure, community detection can be categorized into non-overlapping community detection and overlapping community detection. In this paper, a novel analysis method has been introduced, which analyze the community in a new angle.

The work is supported by National Natural Science Foundation of China (No. 91124002, No. 61372191), National Basic Research Program (973 Program) of China (No. 2013CB329601), and China Postdoctoral Science Foundation Program (No. 2012M520114, No. 2013T60037). Thesis projects aim to study the running mechanisms of social networks, including (1) structural Properties and Evolving laws, (2) Crowds and their interaction behavior and (3) Information and its dissemination. Many papers have been published in respectable international conferences, such as WWW, SIGKDD, SIAM DM etc.