doi:10.3969/j.issn.1000-436x.2014.11.002

基于胜任力模型的社交网络意见领袖识别方法

陈波[1]，唐相艳[1]，于泠[2]，刘亚尚[1]
(1. 南京师范大学 计算机科学与技术学院，江苏 南京 210023；2. 江苏省大规模复杂系统数值模拟重点实验室，江苏 南京 210023)

摘 要：提出了社交网络意见领袖的胜任力模型，模型中包括社交网络意见领袖所应具有的信息生产、信息传播以及信息影响 3 大能力要素及各个显性和隐性行为指标。根据胜任力模型，将社交网络用户划分为普通大众、活跃分子、主题意见领袖和网络意见领袖等 4 类，设计了意见领袖的层次筛选流程和筛选实验系统框架。以采集到的新浪微博数据为例，通过整合主题分类工具 MALLET 及多种社交网络分析工具，再采用 Python 语言针对特定功能编程，实现了意见领袖的识别。实验验证了识别模型的有效性和层次筛选方法的可行性。

关键词：胜任力；意见领袖；社交网络；社会计算
中图分类号：TP315 **文献标识码**：A **文章编号**：1000-436X(2014)11-0012-11

Identifying method for opinion leaders in social network based on competency model

CHEN Bo[1], TANG Xiang-yan[1], YU Ling[2], LIU Ya-shang[1]
(1. School of Computer Science, Nanjing Normal University, Nanjing 210023, China;
2. Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, Nanjing 210023, China)

Abstract: A competency model of opinion leaders in social networks is presented, which includes three capacity factors, some explicit and implicit behavioral indicators. These characteristics of information production, information dissemination and information influence which opinion leaders should have. Based on the proposed competency model, social network users into four categories which divided: the general public, the activists, the topic opinion leaders and network opinion leaders. Then a hierarchical filtering process is designed and a system framework of filtering experimental for identifying opinion leaders is introduced. It takes the data extracted from the Sina microblog as input, then integrates some social network analysis tools such as MALLET, and develops the programs for specific functions with Python. Eventually, the opinion leaders of microblog can be identified. The experiment results prove the effectiveness of present competency model for opinion leaders and the feasibility of hierarchical filtering process.

Key words: competency; opinion leaders; social network; social computing

1 引言

社交网络，即社交网络服务(SNS, social networking services)，是指帮助人们建立社会性网络的互联网应用服务[1]。社交网络由于其具有的开放性、交互性、虚拟性在系统推荐、社会信息安全及知识共享等领域有着广泛的应用前景。然而，社交网络在这些领域中的应用和发展面临一系列新的问题。例如在系统推荐领域，面对处于混沌状态的，海量无序、复杂数据给企业利用社交媒体进行社会化推荐造成的困难[2]，如何帮助信息生产者针对正确的用户进行精准推荐、帮助信息消费者迅速定位自己所需信息。在社会信息安全领域，面对社交网络中信息密度虽低，但是用户关系泛在，传播途径泛在，

收稿日期：2014-08-14；**修回日期**：2014-10-22
基金项目：江苏省教育科学“十二五”规划重点基金资助项目（B-a/2013/01/013）；江苏省教育科学“十二五”规划基金资助项目（D/2013/01/002）
Foundation Items: The Major Program of the 12th Five Years Education Science Plans of Jiangsu Province (B-a/2013/01/013); The Program of the 12th Five Years Education Science Plans of Jiangsu Province (D/2013/01/002)

负面信息在短时间内大范围传播，呈现蝴蝶效应，影响社会稳定的问题[3]，如何针对网络舆情进行有效监管，以保证国家安全和社会稳定。在知识共享领域，面对资源共享和管理的难题[4]，以及社交网络中人际交流中广泛存在的弱连接[5]，如何高效率、低成本地获取有价值信息、实现资源共享，以弥补个人或团体资源的不足。

发现和培养意见领袖是解决这一系列问题的一种重要途径。“意见领袖(opinion leader)”这一概念最早由 Lazarsfield 在 20 世纪 40 年代提出[6]，是指在人际传播中经常为他人提供信息，同时对他人施加影响的“活跃分子”，他们在大众传播效果的形成过程中起着重要的中介或过滤作用，并由他们将信息扩散给受众。后期的学者们对于意见领袖这一概念做出了很多阐释，但是归结起来都强调了意见领袖在信息传播过程中的行为和对传播效果的影响力[7]。

在系统推荐领域，意见领袖作为信息生产者与信息消费者之间的“桥”，可以起到很好的桥接作用，针对特定群体，通过意见领袖传递特定信息，以影响大众舆论和购买决策[8]，实现生产者与消费者双方价值的双赢。在社会信息安全领域，借助意见领袖的力量，可以对网络舆情进行有效监控，促进正能量传播、抑制负面信息传递，确保社会信息安全，意见领袖已经被证明在政治参与及维持长期的政治活动中具有显著的影响力[9]。在知识共享领域，意见领袖可以跨越区域边界，起到中介过滤、控制及桥接作用，更好地实现资源共享[10]。

然而，目前对于社交网络意见领袖的识别工作缺乏一定的理论基础[11]，识别特征指标的设定、识别方法以及识别工具的应用开发还存在很多局限性。为此，本文从胜任力（competence）理论这一新的视角来深入探讨社交网络意见领袖的识别，分析目前社交网络意见领袖识别研究中存在的问题，提出了基于胜任力模型的意见领袖识别方法并进行了实证。

2　相关工作

意见领袖的识别包括 3 个方面的工作：确定识别指标、选择识别方法和运用识别工具，其中确定识别指标是整个工作的基础。文献[12]依据信息传播载体将意见领袖识别技术划分为 3 个发展阶段：传统意见领袖、网络意见领袖和社交网络意见领袖，分析了这 3 个阶段的主要工作及存在的问题。本文用列表形式加以简要说明，如表 1 所示。清华大学刘建明教授在主编的《宣传舆论学大辞典》中，将意见领袖及其所涉及的理论归类于“传播过程和传播效果”，因而本文将意见领袖识别指标分为信息生产（包括过滤、加工、解读）、信息传播过程和信息传播效果 3 个部分。

传统意见领袖是指在政治投票、市场营销、流行时尚、公共事件等借助传统媒体发挥人际影响力的人物。传统意见领袖由于所处的地位高，人际渠道广，处于信息源的上端，容易获得信息，同时由于其通常享有较高威望，在信息的传播中可信度高，因而具有影响力。不过，传统意见领袖的识别指标[6]过于主观，有的难以测量，如威望和影响力。这一时期的识别方法建立在成员相互熟悉的前提下，主要用于小群体研究，不适用于网络环境下大数据量的计算。

借助于网络这个重要的传播与共享平台，信息能够从信息位较高的用户快速流向信息位较低的

表 1　　**意见领袖识别技术**

意见领袖	识别指标			识别方法		识别工具
	信息生产	信息传播过程	信息传播效果	定性	定量	
传统意见领袖	地位；威望	社会交往行为	影响力	自我报告法；观察法	社会计量法；简单统计测量法	意见领袖量表、调查问卷
网络意见领袖	威望度（注册时长）	发帖、回帖数量；平均内容长度；在论坛中的活动时间	被回帖数量，回帖长度；帖子被浏览次数回复者的分散程度响应值、认同值（论坛中帖子获得支持和反对数）	自我报告法；观察法	社会计量法；基于影响力扩散模型；基于聚类	SPSS 等
社交网络意见领袖	威望度（注册时长）	点入度；点出度；度数中心度、中间中心度；网络有效规模、网络约束系数	核心—边缘模型、影响力系数	自我报告法；观察法	基于聚类、基于 PageRank、HITS 算法、社会网络分析法等	UCINET 等

用户，网络环境下的意见领袖由此发挥影响力。网络意见领袖的识别指标在关注用户发帖数量的同时开始考虑信息传播影响力的定量描述，不过也仅限于统计发帖长度、回帖长度、对其发言进行回复的回复者的分散程度以及支持和反对数等[13~15]，对于认同度的度量较主观。仍缺乏对用户所生产信息质量的评价指标，仅通过用户 ID 注册时长描述用户的威望不够准确。

随着社交网络的兴起，意见领袖也显现出了新的特质，社交网络环境中的信息传播不仅更快，而且内容形式更加丰富，用户之间的联系更加广泛和紧密。在社交网络中，不仅存在以微博、帖子等形式的“信息流”[16]，而且还存在以用户为中心的“用户流”，以及基于用户与用户之间关系的“影响力流”。因此，目前针对社交网络意见领袖识别技术在指标设定、识别方法和识别工具的应用上还存在不少问题。

现有对于社交网络意见领袖识别指标的设定缺乏一定的理论基础，没能从意见领袖这些个体的知识、特质、社会角色等方面构建对其的识别特征。现有社交网络意见领袖的识别指标一定程度上考虑了社交网络的交互性、引入了多样化的指标[13,16,17]，但是，其选取的指标不够系统化的问题依然突出。例如，更多的研究集中在意见领袖在社交网络中连接关系的统计上；现有指标[18,19]注重个体间的关系，但是对于个体所生产信息的质量仍然缺乏有效度量。另外，特征指标的选取[7,20]忽略了意见领袖在信息的传播过程中是否可以作为中介者，即起到中介过滤、控制及联络各个小团体的作用。社交网络最大的特点是互动性，意见领袖不仅可以在一个小团体网络中具有“呼风唤雨”的作用，还应当在不同的信息流中，丰富网络的信息与知识，起到中介过滤、控制及桥接的作用，而这也正是发现社交网络意见领袖的核心意义所在。

社交网络的识别方法大多数从属性矩阵数据出发，没有很好地体现社交网络的整体性与互动性。属性矩阵数据更关注的是个体特性，从个体角度进行分析，没有从整体的角度进行分析，不能够很好地体现网络的整体性和互动性。例如应当可以考虑回复数和阅读数之间的比例，以表征该贴的反响程度。关系矩阵数据能够从原本的社会关系结构中分析出个体所处的结构、位置、关系及互动模式。若将属性矩阵与关系矩阵结合研究，则既可以关注个体的属性，又可以注重个体间的互动关系，即整体网络结构。

此外，目前研究中构建的社交网络关系图大多数是无向无权图或无向带权图，没有很好地体现社交网络的网络特性。因为无向图不能够很好地体现用户之间的互动方向关系，无权图也不能够很好地体现用户之间互动的亲密程度。若采用有向带权社交网络关系图，则可以很好地体现社交网络用户之间的互动方向和互动亲密程度。同时，已有的研究对社交网络用户发布的内容研究过少，用户发布的内容是用户情感的体现。

已有社交网络意见领袖识别方法的实现大都是基于单一的社交网络分析工具。目前的社交网络分析工具使用便捷，但是功能略显单一，不能满足用户的特定需求，不方便进行相应的科学实验。整合多种社交网络分析工具，再加以针对特定功能的编程实现，则可以避免单一社交网络工具的局限性，也可以满足实验中的定制功能需求。

本文针对上述已有研究中存在的问题展开研究。社交网络意见领袖是指在社交网络中，具有高的专业性、创新性、专注度（信息生产能力），乐于主动贡献内容、分享数据（信息传播能力），具有较高关注度、认可度，并且能对其他用户的互动起中介过滤、控制及桥接作用，影响舆论导向的活跃分子（信息影响能力）。基于这样的认识，本文采用管理学中提出的并已经得到广泛应用的胜任力模型理论，提出了意见领袖识别方法。

3 基于胜任力的社交网络中意见领袖识别

3.1 胜任力模型

“胜任力”这一概念最早由美国著名心理学家 McClelland 于 1973 年提出。他将胜任力定义为与工作或工作绩效或生活中其他重要成果直接相似或相联系的知识、技能、能力、特质或动机[21]。1994 年，他和 Spencer 进一步将这个概念明确为能将某一工作（或组织、文化）中高绩效者与一般绩效者区分开来的，可以通过可信方式度量出来的知识、技能、社会角色、自我概念、特质和动机等可识别的行为技能和个人特征[22]。

我国学者在研究和应用“胜任力”理论时将其阐释成“胜任能力”、“胜任特征”或“胜任素质”等[11,23]。胜任力的概念包含着对于任务、岗位或职

务要求“胜任”的含义，是综合才能的体现。胜任力的概念在管理领域中受到普遍重视，学者们为制定选拔和任用人才的有效测评指标而研究构建胜任力模型[24]。

胜任力模型是指达到某一绩效目标，角色所需具备一系列不同胜任力要素的总和。胜任力模型通常以 McClelland 给出的胜任力冰山模型为理论基础。该模型将人员个体素质的不同表现形式划分为显性的“冰山水上部分”和隐性的“冰山水下部分”。冰山水上部分包括知识和技能，而冰山水下部分包括社会角色、自我形象、特质和动机。显性胜任力是对胜任者的基本能力要求，隐性胜任力是深层次的潜在能力，是区分高绩效者与一般绩效者的关键因素。

3.2　意见领袖识别的胜任力模型

社交网络环境中，存在着形形色色的用户，意见领袖通常是生活在现实生活中的个体，自然具有区别于一般个体的一些特定的心理和行为特征。因此，可以从胜任力理论这一新的视角来深入探讨社交网络意见领袖的识别特征。

本文基于胜任力理论提出用于精准识别社交网络意见领袖的胜任力模型。模型由 3 个基本胜任力要素组成，如表 2 所示。表中给出了胜任力要素的名称、胜任力要素的定义（即界定胜任力的关键要素的含义）和逐层细化的行为指标（反映胜任力行为表现的差异）。

3.2.1　活跃性

活跃性直接体现了意见领袖在信息传播过程中的活跃程度。用户 u 活跃性的计算式为 3 项的加权和。

$$Activity(u)=\alpha AT(u)+\beta ADP(u)+\gamma ADR(u) \quad (1)$$

其中，α、β、γ 是可以调节的参数，$AT(u)$表示用户关注朋友数目，关注朋友数目越多，说明用户在社交网络活动意愿越强烈；$ADP(u)$表示用户平均每周发文数目，$ADR(u)$表示用户平均每周回复其他文档的数量。

3.2.2　中心性

活跃性侧重反映用户在整个社交网络信息传播过程中的主动活跃程度，而中心性则反映用户的信息生产能力，即用户是否具有很高的受关注度，是否可以影响舆论导向。中心性按主题进行计算，本文借助文档主题生成模型（LDA, latent dirichlet allocation）对用户的发文进行主题分类，然后计算各个主题下的用户中心性指标值，包括影响力、专业性、创新性和专注度 4 个行为测度指标。

LDA 是 Blei 等于 2003 年提出的一个 3 层概率生成模型，是一种非监督机器学习技术[25]。本文借助 LDA 主题模型通过“文档-词语”矩阵进行训练，学习出“文档-主题”、“主题-词语”2 个矩阵，进而就可以知道每篇文档的主题分布，计算中心性的各个行为指标。

1) 影响力

影响力主要基于用户的粉丝数及被回复 2 种行为信息，被回复也包括部分社交网络中的被转发、被“赞”等行为。本文将影响力量化为受众度、反响度、扩散度的加权和。

受众度表示用户 u 发表的某主题 t 文档在社区的知名度，用阅读过此类主题文档的用户数量 $read_u$ 除以网络社区总用户数量 all，计算式为

$$cov_t(u)=read_u/all \quad (2)$$

反响度表示用户文档在社区的反响程度，用回复过此文档的用户数量 $response_u$ 除以阅读过此文档的用户数量 $read_u$，计算式为

$$resp_t(u)=response_u/read_u \quad (3)$$

表 2　社交网络意见领袖胜任力素质模型

胜任力要素	要素定义	行为指标
活跃性（显性）	在信息传播过程中的活跃程度	关注朋友数 平均时间的发文数 平均时间的回文数
中心性（隐性）	所生产信息的质量和效果	影响力：受众度、反响度、扩散度 专业性：主题偏离度 创新性：原创度、文档相似度 专注度：特定主题的发文、回文和转发的比例
中介性（隐性）	信息传播过程中的中介过滤、控制等作用	中介中心度

其中，t 表示特定主题，u 表示特定用户。

扩散度表示文档在社区的扩散快慢，如果一个文本信息被一个粉丝转发，而此粉丝拥有的粉丝数也很多，那么这个文本信息就会在短时间内被更多的人看到。用户的扩散度是每个转发用户的扩散力求和。扩散度计算式为

$$diff_t(u)=\sum_v \frac{1}{forward(u)} fans(v) \tag{4}$$

其中，$forward(u)$ 表示用户 u 的文本信息被转发的数量，$fans(v)$ 表示转发用户 v 的粉丝数。

用户 u 影响力的计算式为该用户关于主题 t 的文档的受众度、反响度、扩散度的加权和。

$$Influence_t(u)=\delta cov_t(u)+\varepsilon resp_t(u)+\chi diff_t(u) \tag{5}$$

其中，δ、ε、χ 是可以调节的参数。

2) 专业性

专业性主要基于用户动态行为信息和行为内容信息进行分析。用户在某个主题发表的信息越多，包含的专业术语越多，表明用户对这个主题越感兴趣、越专业。在此，使用逼近理想解法(TOPSIS, technique for order preference by similarity to an ideal solution)[26]来测量用户整体专业性。

根据 LDA 得出的“文档—主题”矩阵将每篇文档中具有最高概率 d_t 的主题作为文档的主题，把用户发表的关于某一主题的所有文档作为一个集合 $D_{t,u}$，集合中最大值定义为 $d_{t,u}^+$，最小值定义为 $d_{t,u}^-$，这里 t 表示特定主题，u 表示特定用户。定义每篇文档的 d_t 与 $d_{t,u}^+$、$d_{t,u}^-$ 的欧氏距离为 $S_{t,u}^+$、$S_{t,u}^-$，分别表示更贴近主题还是更偏离主题[7]。

用户专业性的计算式为

$$Expertise_t(u)=S_{t,u}^+/(S_{t,u}^+ + S_{t,u}^-) \tag{6}$$

$S_{t,u}^+$、$S_{t,u}^-$ 的计算如下

$$d_t = \text{文档中最高概率的主题} \tag{7}$$

$$d_{t,u}^+ = \max(D_{t,u}) \tag{8}$$

$$d_{t,u}^- = \min(D_{t,u}) \tag{9}$$

$$S_{t,u}^+ = \sqrt{\sum_{d_t \in D_{t,u}} \left(d_{t,u}^+ - d_t\right)^2} \tag{10}$$

$$S_{t,u}^- = \sqrt{\sum_{d_t \in D_{t,u}} \left(d_t - d_{t,u}^-\right)^2} \tag{11}$$

3) 创新性

用户的创新性主要基于行为内容信息分析，从原创度和文档相似性 2 个角度考虑。

原创度表示用户发表的所有文档中，原创文档所占比例的多少。原创度越高，此用户的创新性就越高，原创度的计算式为

$$Creativity_t(u)=1-forward_{t,u}/Sum_{t,u} \tag{12}$$

其中，$forward_{t,u}$ 表示用户 u 关于主题 t 转发的文档数量。

文档相似性主要是计算特定主题内文档之间的相似性。首先，每篇文档表示为一个词频和逆向文本频率（TF-IDF, term frequency-inverse document frequency）权重向量，然后使用余弦的方法计算特定主题之间文档的相似性，选择最大值作为此用户的相似性，记为 $Similarity_t(u)$。

用户创新性的计算式为

$$Noverty_t(u)=Creativity_t(u)\frac{1}{Similarity_t(u)} \tag{13}$$

4) 专注度

专注度用特定主题的发文、回文、转发数占用户总发文、回文、转发数的比例来表示，其值越高，说明用户对主题领域越关注。用 $Focuse_t(u)$ 表示用户 u 在主题 t 的专注度，$Post_{t,u}$ 表示用户 u 对主题 t 发表的所有发文数量，$Reply_{t,u}$ 表示用户 u 关于 t 主题回复的所有回文数量，$Forward_{t,u}$ 表示用户 u 关于主题 t 转发的所有发文数量，关注度的计算式为

$$Focuse_t(u)=\kappa Post_{t,u}\Big/\sum_{t=1}^{n} Post_{t,u}+\eta Reply_{t,u}\Big/\sum_{t=1}^{n} Reply_{t,u}+\lambda Forward_{t,u}\Big/\sum_{t=1}^{n} Forward_{t,u} \tag{14}$$

其中，κ、η、λ 是可调节的参数。

最终，影响力、专业性、创新性、专注度共同构成了用户的中心性指标，其计算式为

$$Centrality_t(u)=\zeta Influence_t(u)+\sigma Expertise_t(u)+\tau Noverty_t(u)+\varphi Focuse_t(u) \tag{15}$$

ζ、σ、τ、φ 这 4 个参数通过结构方程模型（SEM, structural equation modeling）计算，并进行归一化处理后求得。

3.2.3 中介性

中介性表示意见领袖在信息传播过程中作为

中介者，起到中介过滤、控制及联络各个小团体作用的度量。

中介性主要借助结构洞理论测量。结构洞的测量最常用的是Burt代表的结构约束算法和Freeman代表的中介中心度算法，研究实验表明 Freeman 代表的中介中心度算法更为有效[27]。Freeman 认为，如果一个行动者处在许多交往路径上，这个人可以通过中介过滤或控制信息而影响群体。中介性测量的是一个点在多大程度上位于图中其他点的中间。因此，本文采用中介中心度作为中介性进行结构洞的测量[28]。

计算各主题意见领袖的中介性指标值之前，首先需要利用已有数据构建社交网络关系图。本文构建的社交网络关系图为有向带权图，有向带权图更贴合现实社交网络的结构及信息的流向。在有向带权社交网络图 $G=(V,E,W)$中，$V=\{v_1,v_2,\cdots,v_n\}$ 是节点的集合，$E=\{e_1,e_2,\cdots,e_m\}$是节点之间边的集合，$W=\{w_{ij}>0\,|\,i,j=1,2,\cdots,n\}$ 是边的权重集合。在社交网络图 G 中，节点表示社交网络用户，边表示用户之间的互动关系。边的方向表示用户互动的方向，边的权重大小表示用户之间的互动次数，互动次数越高，则相应的权重值越高。

中介中心度的计算步骤如下。

步骤 1　计算点 i 能够控制 j 和 k 这 2 点的交往能力为

$$b_{jk}(i)=g_{jk}(i)/g_{jk} \tag{16}$$

其中，g_{jk} 表示点 j 和 k 之间存在的捷径数目，$g_{jk}(i)$ 表示点 j 和 k 之间存在的经过点 i 的捷径数目。

步骤 2　计算点 i 相对于图中所有点对的中间度的和，然后把图中点 i 相应与图中所有点对的中间度加在一起，得到该点的绝对中介中心度(absolute betweenness centrality)。

$$C_{ABi}=\sum_{j}^{n}\sum_{k}^{n}b_{jk}(i),\ j\neq k\neq i\text{且}j<k \tag{17}$$

步骤 3　计算点 i 的相对中介中心度(relative betweenness centrality)

$$C_{RBi}=\frac{2C_{ABi}}{n^2-3n+2} \tag{18}$$

3.3　意见领袖层次筛选流程

依据胜任力模型将社交网络中的用户分为普通大众、活跃分子、主题意见领袖和网络意见领袖 4 类。本节给出了意见领袖的层次筛选流程，如图 1 所示。

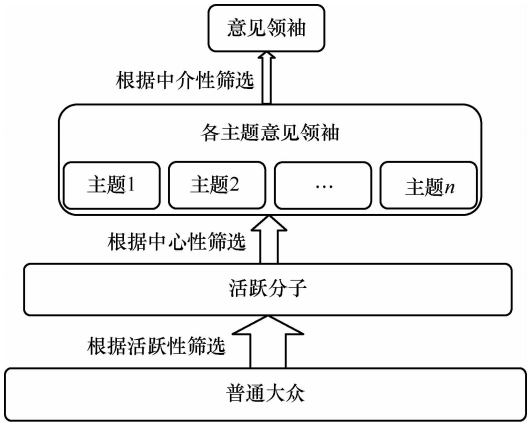

图 1　意见领袖识别层次筛选流程

1) Identify_Activists 算法

该算法根据用户的关注数、发表文档数量、回复文档数量和时间信息，计算出用户的活跃性指标值，活跃性值在阈值 k 之上的作为活跃分子（算法描述略）。

2) Identify_TopicLeaders 算法

在已有的活跃分子发文中，利用 LDA 主题模型对社区所有活跃分子的信息按主题分类。在各主题内部，用该算法计算出用户中心性指标值，中心性值在阈值 k 之上的作为各主题的意见领袖（算法描述略）。

3) Identify_OpinionLeaders 算法

该算法主要是利用结构洞理论，计算出各主题意见领袖的中介性指标，取主题意见领袖中中介性值在阈值 k 之上的作为整个社区的意见领袖。算法描述如下。

OpinionLeaders=∅
Construct graph $g=(V,E,W)$ based on topic
$V=\varnothing$, $E=\varnothing$
for each $id\in$ microblogger
$V\leftarrow id$
if id has relations with forward or comment $id1$
if $<id1,id>\in E$　modify $w(id1,id)$
else $E\leftarrow <id1,id>$, $w(id1,id)=1$
endif
　　endif
　endfor
end Construct
for each $u\in V$
if $u\in$ TopicLeaders
compute Intermediary(u)

endif
if Intermediary(u)>k
OpinionLeaders←u
endif
endfor

4 实证研究

4.1 实验系统框架

本节首先给出社交网络意见领袖识别实验系统框架，接着介绍实验步骤，最后对实验结果进行分析并与相关工作进行了比较。

实验系统框架各步骤主要工作如下。

1) 数据获取。数据获取目前最常用的有2种方法：一是通过爬虫程序获取数据，可以自己编写爬虫程序，也可以利用开源爬虫，如 heritrix、MetaSeeker(GooSeeker)等。二是利用社交网络开放平台提供的应用程序接口 API 进行相关内容的抓取。通过 API，第三方可以按照自己的需求获取不同的数据或者开发不同的应用程序，为此本实验选择新浪微博开放平台获取所需数据。

2) 信息提取。信息提取包括信息定位和信息存储2部分。信息定位主要是从网页中提取与研究相关的信息。通过 API 获取的数据采用 JSON(Java script object notation)格式，这是一种轻量级的数据交换格式，其数据格式简单、应用广泛，可以直接定位相关信息。信息存储则是把定位到的相关信息写入数据库中，以便于后续处理，本实验采用 mySQL 数据库。

3) 社交网络关系图的构建。根据数据库中用户之间的交互关系构建用户交互图。可以直接根据数据库中用户之间的回复关系构建社交网络图，也可以先构建用户之间的关系型矩阵，再把关系型矩阵转换成社交网络图，本实验通过程序设计语言编程实现。此部分还需进行文本分词、主题分类等工作。本实验中的文本分词利用 Python 中文分词组件 jieba 完成[29]；主题分类根据 LDA 主题模型的原理，利用 MALLET(machine learning for language toolkit)[30]软件完成。

4) 意见领袖识别。意见领袖识别是实验框架中最核心的部分，利用 3.3 节给出的层次筛选流程完成。前述的3个筛选算法用 Python 语言编程实现。

5) 结果可视化。社交网络可视化工具可以方便、简单的实现相应数据可视化，但是，它不能根据用户需求灵活地实现定制信息的可视化。本文采用 Python 语言针对特定功能进行编程，可以实现不同的可视化需求，且图形效果也很美观。

4.2 实验步骤及结果

4.2.1 数据获取与信息提取

本文利用新浪微博开放平台提供的API接口函数采集数据。主要接口函数包括：关系读取接口 friendships/friends（获取用户的关注列表）；用户读取接口 users/show（获取用户信息）；微博读取接口 statuses/user_timeline（获取用户发布的微博）和 statuses/repost_timeline（返回一条原创微博的最新转发微博）；评论接口 comments/show（获取某条微博的评论列表）等。获取到 2014 年 1 月 1 日至 2014 年 2 月 15 日期间 435 位用户在新浪微博的发表微博、回复微博等情况，共 7 089 条微博。

数据保存至 MySQL 数据库 microblogsina 中的 2 张表内：microblog_user、microblogger。microblog_user 表存放用户的基本信息，主要字段有 sina_id、nickname、fansnumber、attentions、post_all、comment_all、registerdays，分别表示用户 ID、用户名、粉丝数、关注数、发文数、回文数、注册日期。microblogger 表存放用户发布微博的基本信息，主要字段有 sina_id、txt_id、post_date、forward、comment、read_count、comment_count、forward_count、contents，分别表示用户 ID、文本编号、发文日期、是否转发、是否评论、阅读数、评论数、转发数、文本内容。

4.2.2 意见领袖识别

根据获取的数据，基于意见领袖胜任力模型中的指标和层次化筛选流程进行筛选，各筛选算法用 Python 语言编程实现。

1) 根据活跃性筛选出社交网络的活跃分子

根据活跃性指标计算式(1)，α、β、γ分别取值为 0.4、0.3、0.3，计算出社区所有用户的活跃性指标，阈值 k 设为 100，最终得到 216 位活跃分子。

2) 根据中心性从活跃分子中筛选出主题意见领袖

根据中心性筛选主题意见领袖是通过对活跃分子微博信息主题分类，在各个不同主题内部根据计算式(15)计算出活跃分子中心性值，发现各主题意见领袖。

本文分为 6 个主题，具体内容如表 3 所示。

表 3　　主题及关键词

主题	关键词
主题 0：感悟	人生、快乐、思考、故事、灰飞烟灭
主题 1：公考	公考、公告、行测、申论、公开课
主题 2：娱乐	编剧、原创、品牌、摄影师、主持
主题 3：美食	宵夜、美味、蔬菜、美食、年夜饭
主题 4：学术	博士、技术人才、学者、老师、IT
主题 5：工作	程序员、统计、论坛、数据、开发

主题分类完成后，就可以计算中心性的行为指标值，δ、ε、χ、κ、η、λ 分别设定为 0.4、0.4、0.2、0.3、0.4、0.3。4 个行为指标计算完成后，借助结构方程模型 AMOS 软件得到各个权重并归一化处理[7]，从而得出各活跃分子的中心性值。本文采用软件 AMOS17.0，以上述计算的影响力、专业性、创新性、专注度 4 个行为指标作为自变量，人工评测中心性值作为因变量，建立结构方程模型并进行路径分析，如图 2 所示。

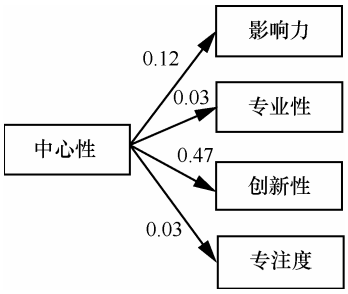

图 2　中心性结构方程模型路径

对中心性结构方程模型路径分析得到的路径系数，进行归一化处理，其计算过程如下。

$$\zeta = 0.12/(0.12+0.03+0.47+0.03) \approx 0.18$$

$$\sigma = 0.03/(0.12+0.03+0.47+0.03) \approx 0.05$$

$$\tau = 0.47/(0.12+0.03+0.47+0.03) \approx 0.72$$

$$\varphi = 0.03/(0.12+0.03+0.47+0.03) \approx 0.05$$

可以看出，影响力、专业性、创新性、专注度系数分别为 0.18、0.05、0.72、0.05，进而计算出各个活跃分子的中心性值。本文把中心性值大于阈值 2.5 的作为每个主题的主题意见领袖，共得到 66 位用户。

3) 根据中介性发现真正的意见领袖

在主题意见领袖中，把中介性值大于阈值 k 的作为真正的意见领袖。中介性主要是根据中介性公式(16)~式(18)计算主题意见领袖的中介性值，其值在阈值 0 之上的即为社区的意见领袖。本实验共得到 13 位意见领袖，如表 4 所示。

作者还借助 Python 语言编程，通过调用 NetworkX 及 Matplotlib 包，按主题构建需要的社交网络关系图，并实现了实验结果的可视化，限于篇幅，图略。

4.3　结果分析与工作比较

目前，针对网络意见领袖的研究大都可归于侧重于关系结构和侧重于内容 2 大类。为了表明本文所提出的意见领袖胜任力识别模型的有效性和层次筛选方法的可行性，本文分别选取了这 2 类中具有代表性的 PageRank 算法、HITS 算法和影响力扩散概率模型 3 种具体的方法进行了比较实验。表 4 列出了 4 种方法选出的意见领袖。

表 4　　比较实验结果

方法	前 13 名意见领袖
PageRank	2543873754、1657809832、1952378163、3314238955、1117617663、1497587753、1774151167、1055769364、1965389715、1035850842、3256814345、1799576054、2802121021
HITS	1670800485、3179781755、1745984651、1713926427、2001348433、2646013373、2083844833、1824294563、1580353657、1736329970、1657809832、1652526230、1799576054
影响力扩散概率模型	1903721143、2181434512、1836678361、2892905951、1541603965、3966555126、2629632621、2509471860、1789356785、1812882783、1999536911、2041028560、1963752797
本文	1909378971、2543873754、1799576054、2935818920、1847022404、2697613543、2521409667、1812882783、3697737621、1998679021、2534483191、1912607653、2179052747

4.3.1　与基于 PageRank 算法识别方法的比较

利用已有的 PageRank 算法[18,19]识别出的前 13 位意见领袖如表 4 所示。PageRank 算法识别出的最核心的意见领袖是用户“2543873754”。然而，通过本文模型计算此用户的活跃性值为 2.47，中心性值为 2.1。该用户一定程度上可以看做社区意见领袖，但不能认为是最核心的意见领袖。与本文实验筛选出的意见领袖“1909378971”相比，用户“2543873754”的活跃性还可以；然而，中心性筛选时，根据图 3 和图 4 截取的该用户实际发文内容可以看出，用户“2543873754”发文短小，实质内容同样较少，而用户“1909378971”发文较长，且主题内容丰富，具有很强的吸引力。

由此可见，单纯利用 PageRank 算法通过计算网络的连接关系得出的意见领袖在一定程度上带有片面性，仅关注了用户之间的链接关系，没有充

sina_id	contents	post_date	forward	comment	read_count	comment_cc	forward_c	detail
2674457074	2674457074-1.txt	2014-02-13 00:13:00	0	2802121021	0	0	0	这么进吗？[思考][思考][吃惊]我下载个试试
2034854711	2034854711-1.txt	2014-02-13 00:17:00	0	2802121021	0	0	0	26！
2543873754	2543873754-1.txt	2014-02-14 23:16:00	0	0	0	2	0	宋小宝是不是那个谁的徒弟
1854251524	1854251524-1.txt	2014-02-14 23:30:00	0	2543873754	0	0	0	是[捂鼻屎]
3582134937	3582134937-1.txt	2014-02-14 23:30:00	0	2543873754	0	0	0	赵本山？
2543873754	2543873754-2.txt	2014-02-14 19:24:00	0	0	9	3	20	这个啥时候cr：
3956328446	3956328446-1.txt	2014-02-14 19:27:00	0	2543873754	0	0	0	[思考]应该就是最近吧 华丽丽的中分
1222018477	1222018477-1.txt	2014-02-14 19:28:00	0	2543873754	0	0	0	像是今天的
2945392284	2945392284-1.txt	2014-02-15 12:28:00	0	2543873754	0	0	0	我也想知道呢！
2543873754	2543873754-3.txt	2014-02-14 19:06:00	2569477485	0	3	2	0	中官说今天录音，应该是不会骗人的 //@小
3151577624	3151577624-1.txt	2014-02-14 19:14:00	0	2543873754	0	0	0	照片好像换点[哈哈]
3033633491	3033633491-1.txt	2014-02-14 20:06:00	0	2543873754	0	0	0	H合体了[抓狂][抓狂][抓狂][抓狂][爱你][爱你
2543873754	2543873754-4.txt	2014-02-14 16:50:00	2599671250	0	3	5	69	主要是好感动先到中官走了一阿
2543873754	2543873754-5.txt	2014-02-14 17:31:00	0	2305961070	0	0	0	留言了吗
2543873754	2543873754-6.txt	2014-02-14 19:17:00	0	2305961070	0	0	0	那就不算

图 3　用户“2543873754”发文内容（阴影部分）

1909378971	1909378971-1.txt	2014-02-02 13:15:56	0	1035850842	0	1	0	电子科大诚邀各路青年才俊下月底聚会成都，一
1035850842	1035850842-6.txt	2014-02-02 12:34:08	0	1985499101	0	0	0	谢谢晓如兄支持
1035850842	1035850842-7.txt	2014-02-02 13:42:48	0	1963752797	0	1	0	可以填会议，你自己制造一个表格，格式可以任
1963752797	1963752797-1.txt	2014-02-02 12:52:48	0	1035850842	0	1	0	支持下家乡的学校！貌似代表性论文只能填期刊
1035850842	1035850842-8.txt	2014-02-02 21:39:38	0	1693138227	0	1	0	申请吧，我们每个申请都会回复，都有惊喜！
1693138227	1693138227-1.txt	2014-02-02 20:15:26	0	1035850842	0	1	0	不是＂具有海外知名大学博士学位，或者具有国
1035850842	1035850842-9.txt	2014-01-05 11:22:04	0	0	10	16	157	最近和亚利桑那州立大学合作发表了人美兴趣动
1035850842	1035850842-10.txt	2014-01-17 09:56:49	0	2350627020	0	1	0	现在好地方卖出去不容易，要花的时间太长，而
2350627020	2350627020-1.txt	2014-01-10 23:21:38	0	1035850842	0	1	0	这文章，价格卖的太低了
1035850842	1035850842-11.txt	2014-02-17 09:57:26	0	1569469735	0	0	0	谢谢常政兄谬赞
1569469735	1569469735-1.txt	2014-01-14 22:58:27	0	1035850842	0	1	0	好文章！要是有中文版的就好了。
1035850842	1035850842-12.txt	2014-01-01 21:49:25	0	0	3	2	4	16篇引用超过100的论文。 论文1+论文2：http:
1909378971	1909378971-2.txt	2014-02-13 17:58:27	0	0	23	5	1	刚收到IFAC录用通知，3年一届的IFAC今年将在
1909378971	1909378971-3.txt	2014-01-30 22:07:34	0	0	11	5	11	大年夜和女儿谈了许多非洲一些国家出现的'资源
1909378971	1909378971-4.txt	2014-01-24 07:21:21	1747250952	0	2	7	9	我在求学阶段最快乐的考试记忆就是考英语，因
2066963092	2066963092-1.txt	2014-01-24 07:35:36	0	1909378971	0	0	0	我做梦常常是一进车间停电了！哈哈，这就叫压

图 4　用户“1909378971”发文内容（阴影部分）

分考虑用户的发文内容，更谈不上对发文质量和效果的考量。

4.3.2　与基于 HITS 算法识别方法的比较

利用已有的 HITS 算法识别意见领袖的方法[18,19]，筛选出的 13 位意见领袖如表 4 所示。HITS 筛选出的核心意见领袖是用户“1670800485”。通过本文模型计算此用户的活跃性指标为 0.567，远达不到本文的活跃性筛选阈值 1.0。同时，根据数据库中原始数据可以得知，此用户的粉丝数为 288，关注数为 125，发表微博数 2 384，显然这些指数相对都是较低的。同时，用户的发文内容长度短，实质内容少，如图 5 所示。实际上，采用 HITS 算法筛选社区意见领袖也有一定的片面性，因为该方法对用户隐性的发文质量也没有考虑。

4.3.3　与基于影响力概率扩散模型识别方法的比较

利用已有的影响力概率扩散模型识别意见领袖方法[20]，筛选出的 13 位意见领袖如表 4 所示。可以看出，该模型筛选出的核心意见领袖是用户“1903721143”。通过本文模型计算此用户的活跃性值为 2.3，中心性值为 6.2。查看该用户资料可知，该用户是“微博女郎”，标签为“时尚、音乐、听歌”，通过查阅微博发现，其原创内容大多是娱乐方面的，这就造成了其在主题 2 的中心性特别高。影响力概率扩散模型侧重内容分析，根据文献[20]的定义，用户的影响力为用户所发全部有效帖子的影响力之和，这就造成这类在单个主题特别突出而在其他主题几乎不涉及的用户被选为意见领袖。这样得到的意见领袖作为某类特定话题的领袖是可

1670800485	1670800485-1.txt	2014-02-07 13:34:43	1592994155	0	0	2	0	胖若两人
1670800485	1670800485-10.txt	2014-01-25 10:35:39	0	0	0	4	0	自由散漫的日子，生活毫无规律，这样下去，人
1670800485	1670800485-11.txt	2014-01-25 13:04:13	0	1807754157	0	0	0	太闲散了，两本书看完了，年底不再买书了。幸
1670800485	1670800485-12.txt	2014-01-26 18:41:45	0	1807754157	0	0	0	哈哈，你快回来
1670800485	1670800485-13.txt	2014-01-23 13:29:04	1653603955	0	0	3	0	哇瞧，我最近的睡姿耶
1670800485	1670800485-14.txt	2014-02-12 15:45:27	0	2053088672	0	1	0	刚把余额宝钱转走就回升了
1670800485	1670800485-15.txt	2014-01-22 22:22:44	0	0	0	2	0	家乡变的我都快不认识了
1670800485	1670800485-16.txt	2014-01-23 09:16:26	0	2053088672	0	0	0	知道呀，入职有填干部表的，操心也没用，天要
1670800485	1670800485-17.txt	2014-02-12 15:51:50	0	1670800485	0	0	0	恐怖袭击[哈哈]
1670800485	1670800485-18.txt	2014-01-21 14:33:39	1642512402	0	0	5	0	真为一些疯狂的行为担忧祖国的未来

图 5　用户“1670800485”发文内容（阴影部分）

行的，但不能作为真正具有桥接作用的意见领袖。

4.3.4　4 种识别方法的信度对比结果

迄今为止，还没有一个公认的意见领袖识别验证评估方法。为了验证 4 种识别方法的有效性，仿效文献[20]给出的“正确率”评价方法，给出意见领袖集合的信度定义。文献[20]中利用人工分析的结果作为正确率的指标，本文认为这样的评价指标人为因素过多，因此改用新浪微博的认证用户比例（V）与微博达人比例（D）作为验证指标，考虑到微博认证用户比之微博达人的影响力大，给出识别出的意见领袖集合的信度为 P=0.6V+0.4D。比较结果如表 5 所示。

表 5　　4 类识别方法的信度对比结果

方法	认证用户比例	微博达人比例	信度 P
PageRank	4/13	2/13	0.246 2
HITS	2/13	5/13	0.246 2
影响力扩散概率模型	3/13	5/13	0.292 3
本文	5/13	6/13	0.415 4

从表 5 中可以看出，本文识别方法的信度要优于其余 3 种。PageRank 和 HITS 算法的信度一致，影响力扩散概率模型略高。从这 4 种方法的比较可以看出，内容是微博的灵魂，分析内容是选择意见领袖必不可少的基础，本文将内容分析与结构分析相结合，效果更佳。

5　结束语

意见领袖的识别已经成为解决系统推荐、社会信息安全、知识共享等领域面临问题的一项重要工作。本文在深入分析意见领袖识别的演进路线及已有工作基础上，应用胜任力理论来深入探讨社交网络意见领袖的识别特征，提出了社交网络意见领袖的胜任力模型。该模型不仅充分考虑了用户的显性能力：信息传播过程中的活跃度，更加注重对用户在信息传播过程中信息生产能力、信息生产质量及效果等隐性能力的综合考量。

根据胜任力理论，将社交网络中的用户划分为普通大众、活跃分子、主题意见领袖和网络意见领袖等 4 类。提出的层次筛选方法，筛选过程更细致。同时充分考虑了用户自身行为、用户发文信息分析及用户之间互动关系等信息，不仅有属性矩阵数据的分析，也有关系型矩阵的分析，更能体现社交网络的互动性和文本语义信息的丰富性。

最后，构建了意见领袖层次筛选实验系统框架，此框架主要包括网页抓取、信息提取、社交网络关系图构建、意见领袖筛选、结果可视化 5 个部分。各个模块灵活衔接，大部分基于程序设计语言实现，这样可以灵活地实现各种定制功能，为今后意见领袖识别的实验工作提供了借鉴。

以采集到的新浪微博数据为例，通过整合主题分类工具 MALLET 及多种社交网络分析工具，再采用 Python 语言针对特定功能进行编程，实现了意见领袖的识别。与已有的侧重于关系结构的 PageRank 算法、HITS 算法以及侧重于内容的影响力扩散概率模型的比较表明，应用胜任力模型识别的意见领袖由于分析了网络结构、链接关系，以及用户行为表现和发文质量，因而筛选结果更符合真实社交网络的情况。

借助本文识别的社交网络意见领袖，将进一步研究如何让其在系统推荐、社会信息安全、知识共享领域发挥“风向标”的作用。

参考文献：

[1] ELLISON N B. Social network sites: definition, history, and scholarship[J]. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230.

[2] 王飞跃, 曾大军, 曹志冬. 网络虚拟社会中非常规安全问题与社会计算方法[J]. 科技导报, 2011, 29(12): 15-22.
WANG F Y, ZENG D J, CAO Z D. Social computing methods for non-traditional security challenges enabled by the social media in cyberspace[J]. Science & Technology Review, 2011, 29(12): 15-22.

[3] 陈波, 于泠, 刘君亭. 泛在媒体环境下的网络舆情传播控制模型[J]. 系统工程理论与实践, 2011, 31(11): 2140-2150.
CHEN B, YU L, LIU J T. Dissemination and control model of internet public opinion in the ubiquitous media environments[J]. Systems Engineering-Theory & Practice, 2011, 31(11): 2140-2150.

[4] 吕刚. 区域信息资源共建共享中的结构洞现象研究[J]. 图书馆理论与实践, 2012, (2): 58-59.
LV G. Study on phenomena of structural holes in the area of information resource sharing[J]. Library Theory and Practice, 2012, (2): 58-59.

[5] 刘广为, 杨雅芬, 张文德. 科技资源共享中“桥”的作用——基于人际网络“结构洞”理论的研究[J]. 图书情报工作, 2009, 53(20): 60-64.
LIU G W, YANG Y F, ZHANG W D. The “bridge” application in sharing of scientific and technological resources—based on the “structural holes” theory of the social network[J]. Library and Information Service, 2009, 53 (20): 60-64.

[6] LAZARSFIELD P F, BERELSON B, GAUSET H. The People’s Choice: How the Votes Makes Up His Mind in a Presidential[M]. New York: Columbia University Press, 1948.

[7] LI Y, MA S, ZHANG Y, *et al*. An improved mix framework for opinion leader identification in online learning communities[J]. Knowl-

edge-Based Systems, 2013, (43):43-51.

[8] CHO Y, HWANG J, LEE D. Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach[J]. Technological Forecasting and Social Change, 2012, (79):97-106.

[9] VAN J L. Political engagement and government informing seeking: increasing role of social media and mobile devices [EB/OL]. http://www.academia.edu/5297330, 2014.

[10] CHOI J H, SCOTT J E. Electronic word of mouth and knowledge sharing on social network sites: a social capital perspective[J]. Journal of theoretical and applied electronic commerce research, 2013, 8(1): 69-92.

[11] 虞鑫, 戴必兵. 醒客工场[EB/OL]. http://www.Thinkerworks.cn,2014.
YU X, DAI B B. Thinkerworks[EB/OL]. http://www.thinkerworks.cn, 2014.

[12] 唐相艳, 于泠, 陈波. 意见领袖筛选方法研究[J]. 情报探索, 2013, 7:6-10.
TANG X Y, YU L, CHEN B. Study on methods of identifying opinion leader[J]. Information Research, 2013, 7:6-10.

[13] 祝帅, 郑小林, 陈德人. 论坛中的意见领袖自动发现算法研究[J]. 系统工程理论与实践, 2011, 31(2): 7-12.
ZHU S, ZHENG X L, CHEN D R. Research of algorithm for automatic opinion leader detection in BBS[J]. Systems Engineering- Theory, 2011, 31(2): 7-12.

[14] 王珏, 曾剑平, 周葆华等. 基于聚类分析的网络论坛意见领袖发现方法[J]. 计算机工程, 2011, 37(5): 44-49.
WANG J, ZENG J P, ZHOU B H, *et al*. Online forum opinion leaders discovering method based on clustering analysis[J]. Computer Engineering, 2011, 37(5): 44-49.

[15] 薛可, 陈晞. BBS 中的“舆论领袖”影响力传播模型研究——以上海交通大学“饮水思源“BBS 为例[J]. 新闻大学, 2010, 4:87-93.
XUE K, CHEN X. Study on the influence propagation model of opinion leader in BBS—the GRATEFUL BBS of SJTU taken for example[J]. Journalism Quarterly, 2010, 4: 87-93.

[16] 丁汉青, 王亚萍. SNS 网络空间中“意见领袖”特征之分析——以豆瓣网为例[J]. 新闻传播研究, 2010, 3: 82-91.
DING H Q, WANG Y P. Analyzing opinion leader attributes in SNS cyberspace: an investigation of douban.com[J]. Journalism & Communication, 2010, 3: 82-91.

[17] NING M, YIJUN L, RUYA T, *et al*. Recognition of Online Opinion Leaders Based on Social Network Analysis[M]. Berlin: Springer Berlin Heidelberg, 2012.

[18] WENG J, LIM E P, JIANG J, *et al*. Twitterrank: finding topic-sensitive influential twitterers[A]. Proceedings of the third ACM International Conference on Web Search and Data Mining[C]. 2010. 261-270.

[19] TANG X, YANG C C. Identifying influential users in an online healthcare social network[A]. Proceedings of IEEE International Conference on Intelligence and Security Informatics[C]. 2010.43-48.

[20] 樊兴华, 赵静, 方滨兴等. 影响力扩散概率模型及其用于意见领袖发现研究[J]. 计算机学报, 2013, 36(2): 360-367.
FAN X H, ZHAO J, FANG B X, *et al*. Influence diffusion probability model and utilizing it to identify network opinion leader[J]. Chinese Journal of Computers,2013,36(2):360-367.

[21] MCCLELLAND C D. Testing for competence rather than for intelligence[J]. American Psychologist, 1973, 28(1):1-24.

[22] SPENCER L M, MCCLELLAND D C, SPENCER S. Competency Assessment Methods: History and State of the Art[M]. Boston: Hay-McBer Research Press, 1994.

[23] 张梅英. 胜任力研究综述[J]. 生产力研究, 2012, 12: 250-252.
ZHANG M Y. Review of competency[J]. Productivity Research, 2012, 12:250-252.

[24] 徐峰. 人力资源绩效管理体系构建:胜任力模型视角[J]. 企业经济, 2012, 1: 68-71.
XU F. Human resource performance management system building: competency model perspective[J]. Enterprise Economy, 2012, 1: 68-71.

[25] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. the Journal of Machine Learning Research, 2003, 3: 993-1022.

[26] TZENG G H, HUANG J J. Multiple Attribute Decision Making: Methods and Applications[M]. Florida: CRC Press, 2011.

[27] 汪丹. 结构洞算法的比较与测评[J]. 现代情报, 2008, 9: 153-156.
WANG D. A comparative study on algorithm of structural holes[J]. Modern Information, 2008, 9: 153-156.

[28] 刘军. 整体网分析讲义[M]. 上海: 格致出版社, 2009.
LIU J. Lectures on Whole Network Approach[M]. Shanghai: Truth & Wisdom Press, 2009.

[29] 开源中国. Python 中文分词组件 jieba [EB/OL]. http://www.oschina.net/p/jieba, 2014.
Open source China. Python Chinese word components jieba[EB/OL]. http://www.oschina.net/p/jieba, 2014.

[30] MALLET. Machine learning for languagE toolkit [EB/OL]. http://mallet.cs.umass.edu, 2014.

作者简介：

陈波（1972-），男，江苏南通人，南京师范大学教授，主要研究方向为移动安全、网络与信息安全。

唐相艳（1988-），女，山东临沂人，南京师范大学硕士生，主要研究方向为信息安全、社会计算。

于泠（1971-），女，江苏金坛人，南京师范大学副教授，主要研究方向为信息安全、社会计算。

刘亚尚（1990-），女，河南郑州人，南京师范大学硕士生，主要研究方向为信息安全、社会计算。