基于时间变化图的网络论坛意见领袖识别算法

徐会杰 蔡皖东 王剑平 陈桂茸

（西北工业大学计算机学院 西安 710072）

摘 要 针对现有意见领袖识别算法难以捕获网络的动态特性这一现状，提出了一个基于时间变化图的网络论坛意见领袖识别算法。该算法将网络论坛的演变描述为一连串静态图，每一幅图代表一个给定时间窗口内用户间的所有交互。依据构造的量化指标识别不同时间窗口内的潜在意见领袖，这些意见领袖然后和其他时间窗口上的意见领袖相匹配以便识别随时间推移的真正意见领袖。实验结果证实了该算法的可行性和有效性。

关键词 时间变化图，网络论坛，意见领袖，时间窗口

中图法分类号 TP393 **文献标识码** A

Identifying Algorithm for Opinion Leaders of Forums Based on Time-varying Graphs

XU Hui-jie CAI Wan-dong WANG Jian-ping CHEN Gui-rong

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Because the existing methods for identifying opinion leaders of the forums are diffcult to capture the dynamic characteristic of network, a algorithm for identifying opinion leaders based on time-varying graphs was proposed. The algorithm represents the evolution of the network as a sequence of static graphs, each of which represents the aggregated interactions over a given time-window. Each potential opinion leader is identified by a quantitative indicator, and these results are then matched each other so that reliable opinion leaders can be identified. Results prove that the proposed algorithm has good validity and feasibility.

Keywords Time-varying graphs, Forum, Opinion leaders, Time-window

1 引言

网络论坛的网络化和交互性使得用户可以在线讨论和散布关于商品以及制造商的各种数据，这些数据提供了依据个人经验和观点并与企业和销售机构相关的大量信息。统计数据显示，网络论坛的大部分用户不经常参与信息的制造与传播，用户做出的决定往往跟随意见领袖。有效地识别网络论坛的意见领袖，通过意见领袖向所在网络的用户投放产品和服务评论而非直接说服用户，可以有效地触发整个网络论坛或社会的影响力，对于减少企业广告费用支出和提高销售效益具有重要的现实意义。

关于网络论坛意见领袖的识别问题，目前国内外研究者已做了广泛研究。Hon Wai Lam 等[1]为了发现在线拍卖网站 eBay 上不同社区中的意见领袖，提出了一个基于社会网络分析的 BuyerRank 模型，该模型根据买家以往的拍卖/购买行为估算他们未来的影响力，并基于他们的影响力对潜在的买家进行排名以协助营销人员作出决策。Xuning Tang 等[2]认为一个具有影响力的用户帖子应具有针对某一主题的较高的内容相似度和较快的回复即时性，由此提出一个包含有用户帖子内容相似性和回复即时性的权值函数，据此探测有影响力的用户。Freimut Bodendorf 等[3]根据提取出的论坛中的用户、观点和关系建立社会网络，借此分析中的中心度、密度、Randic 连接性和接近中心势来探测意见领袖和意见趋势。Amir Afrasiabi Rad 等[4]提出了一种根据社会网络中用户间的交互来探测其中最有影响力用户的方法。该方法通过扑捉用户间的交互频率来估算用户间的关系强度和用户的影响力。上述方法以社会网络分析法为主要手段，通过社会网络分析中的相关量化指标发现论坛中有影响力的用户。面对大量虚拟社区中的海量信息，这些方法相对于利用文本挖掘探测用户话题内容和关系的传统方法来说，具有简洁、高效等优点。不足的是这些方法集中于静态网络的结构和统计学特性方面，并且一般不能捕获网络中用户和用户间交互的动态演变特性，如用户在论坛中的角色和权限会随时间的推移而发生变化等，这将对意见领袖识别的准确度产生影响。

针对上述问题，本文提出了一个基于时间变化图的论坛意见领袖识别算法。该算法将社会网络分析法和时间变化图相结合，弥补了上述识别方法的不足，提高了意见领袖的识别准确度。下面首先构建基于网络论坛网络特征的论坛网络；然后讨论意见领袖的行为特征，并提取出意见领袖的属性指标；进而对意见领袖的属性指标进行数据归一化处理；之后通过对比基于静态网络图的分析结果，验证应用该算法的有效性；最后对全文进行总结展望。

到稿日期：2011-10-19 返修日期：2012-02-17 本文受西北工业大学基础研究基金(JC201149)资助。

徐会杰(1980－)，男，博士生，主要研究方向为网络信息安全与信息对抗，E-mail：xhj004@gmail.com；蔡皖东(1955－)，男，教授，博士生导师，主要研究方向为分布式计算、网络信息安全与对抗。

2 网络论坛的网络特征及描述

网络论坛中用户间的交互首先是以第一个作者提出一个主题(thread),该主题有一个标题(title)并且在该 thread 中是唯一的,然后由其他用户(可以是第一个作者)围绕这个标题通过发一个或多个包含相应内容的帖子(post)展开讨论[5]。图 1 所示为网络论坛 thread 组成图。此时,如果将用户作为节点,一个用户对另外一个用户的回帖视为施加一个影响(如 B 对 A 进行了回复,则 B 对 A 施加了一个影响),则可以构造出如图 2 所示的网络论坛中用户间发生互联的有向网络图。设定在一时间周期内,用户间就某一主题发出一定数量的消息,则可将图 2 转变为如图 3 所示的以传递的消息数作为边的权值的有向网络图。

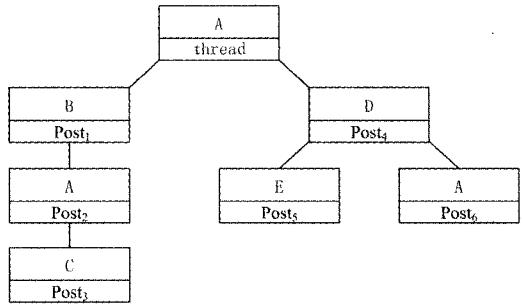

图 1 论坛 thread 组成图

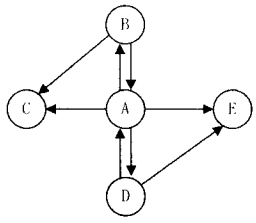

图 2 论坛有向网络图

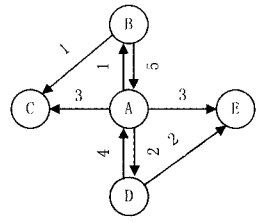

图 3 论坛有向权值网络图

事实上,和其他对象组成的复杂网络一样,基于网络论坛中用户间的交互建立起来的网络也是动态变化的。在论坛的有向权值网络图中,随着时间的推移,节点会不断地加入或离开网络,节点间的边会因此发生搭建或消亡,边的权值以及节点在网络中的角色和权限也会随之发生动态的改变。如图 4 所示,最终的论坛有向权值网络图可能是通过 3 个时间周期 T_1, T_2, T_3 演化而来的。值得注意的是,为了使图保持在一个合理的尺寸,在不影响最终意见领袖识别计算的前提下,一些权值为 0 或接近于 0 的边和节点将被移除(图 4 中无体现,详见第 3 节)。可以通过赋予图中的节点和边一定的生命周期来刻画论坛网络图随时间演变的动态特性。

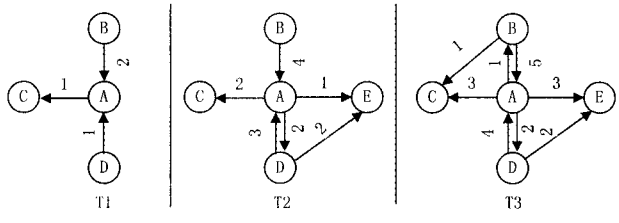

图 4 论坛的动态演变过程

根据以上分析,以网络论坛中的用户作为节点 V,用户间的交互作为边 E,交互的信息数量作为边的权值 W,交互的持续时间 T 作为图的生命周期,图可形式化表示为:$G=(V,E,W,T)$。其中图的生命周期 T 被分割成连续的子区间 $T=[t_0,t_1),[t_1,t_2),\ldots,[t_k,t_{k+1})$,此处每一个区间 $[t_k,t_{k+1})$ 可以用 T_k 表示,则在每一个 T_k 上的论坛网络图可以表示为 $G^k=(V^{[t_k,t_{k+1})},E^{[t_k,t_{k+1})},W^{[t_k,t_{k+1})},T_k)$,每一个时间窗口 T_k 对应的是网络的一个即时快照。所以,一个论坛有向权值网络图可以表示为一连串在各个时间窗口内的图的序列[6],即 $SF(T)=G_0,G_1,\ldots,G_k$。

3 基于动态网络的论坛意见领袖识别

3.1 论坛意见领袖特征分析

意见领袖是在观点的形成与传播中扮演重要角色的人。他们在一个网络中的特殊位置以及交流习惯可以影响其他用户的观点,给那些搜寻信息的用户提供一种导向。研究发现,网络论坛中的意见领袖在行为上一般具有以下两方面特征[2,3,7]:

(1)意见领袖总是与论坛中许多用户存在直接联系;

(2)在一定时间周期内,意见领袖往往非常频繁地直接与论坛中多个用户发生交互,并频繁向多个用户回复信息。

3.2 论坛意见领袖识别

依据网络论坛中意见领袖的行为特征,采用社会网络分析法中的度和聚类指标来进行量化。

度是衡量一个用户直接与其他用户交流的频繁程度。有向网络中节点的度又分为出度(Out-degree)和入度(In-degree)。在本文中,定义出度和入度分别表示一个用户在某一时间周期内发出和接收到的消息数量。如果一个节点具有一个高出度值,表示其指代的用户在网络中创造出比他人更多的内容,所以有更多的机会通过自己的行为去影响他人;如果一个节点具有一个高的入度值和一个非常低的或为 0 的出度值,表示其指代的用户是一个完全不活跃的用户。出度和入度可以用下面的公式表示:

$$D_O(i)=\sum_{j\in N}\vec{e}(i,j)w_{i,j} \tag{1}$$

$$D_I(i)=\sum_{j\in N}\vec{e}(j,i)w_{j,i} \tag{2}$$

式中,i,j 分别代表图中的两个节点,$\vec{e}(i,j)\in E$ 代表从节点 i 到节点 j 的一个有向边,$w_{i,j}\in W$ 代表有向边的权值,N 代表节点 i 的邻接节点集。

聚类是衡量一个用户与一个高度互联的用户集群的亲密程度。在有向网络中,节点的聚类又分为引入聚类(Incoming-clustering)和外出聚类(Outgoing-clustering)。本文中,外出聚类和引入聚类分别表示一个用户在某一时间周期内向集群发出或接收到集群发送来的消息数量。一个节点具有高的外出聚类值,表示该节点指代的用户发出的消息可以快速地在集群内部散播并可以通过集群散播到集群外更大的范围。所以,一个有高外出聚类值的用户具有较高的机会成为意见领袖;同理,如果一个节点具有高的引入聚类值,表示该节点与更多的集群相连接,因此其指代的用户有更多的机会参与了解不同的信息,也即意味着有更高的接受他人意见的机会。外出聚类和引入聚类值可以用下面的公式表示[4,8]:

$$C_O(i)=\frac{\sum_{j\in N}D_O(j)}{D_O(i)\times(D_O(i)-1)} \tag{3}$$

$$C_I(i)=\frac{\sum_{j\in N} D_I(j)}{D_I(i)\times(D_I(i)-1)} \tag{4}$$

式中，j 表示集群中的节点，$\sum_{j\in N} D_O(j)$ 和 $\sum_{j\in N} D_I(j)$ 表示集群中节点间实际存在的边。

通过以上分析可以得出：具有高出度值和高外出聚类值的用户有很大的影响力，也即有很大的可能成为某一时间窗口内的意见领袖。但值得注意的是，一个具有 0 入度且高出度的用户很可能是一个恶意的信息发布者。因此，构造以下等式来量化用户的影响力：

$$Influence(i)=\tanh(D_O(i))*(\alpha D_O(i)+\beta C_O(i)) \tag{5}$$

式中，α,β 为加权值；$tanh(D_O(i))$ 表示出度的双曲正切值，用以表示出度等于(或接近)0 时，用户最终的影响力值将等于(或接近于)0。

在确定上式中 $D_O(i)$ 和 $C_O(i)$ 相对于彼此的重要程度之前，加权值 α,β 是未知的。本文采用 AHP 法求解，根据 Saaty "1-9"标度法构造判断矩阵：

$$R=\begin{bmatrix} A_1/A_1 & A_1/A_2 \\ A_2/A_1 & A_2/A_2 \end{bmatrix}=\begin{bmatrix} 1 & 3 \\ 1/3 & 1 \end{bmatrix} \tag{6}$$

分别计算两行元素的几何平均值后作归一化处理，并计算最后权重，计算公式为：

$$W_i=\frac{W_i'}{\sum_{i=1,\cdots,n} W_i} \tag{7}$$

式中，W_i 为权重，W_i' 为判断矩阵中每行元素的几何平均值。

α 和 β 的计算结果为：

$$\alpha=\frac{\sqrt{1\times 3}}{\sqrt{1\times 3}+\sqrt{1\times\frac{1}{3}}}=\frac{1.7321}{1.7321+0.5774}\approx 0.75$$

$$\beta=\frac{\sqrt{1\times\frac{1}{3}}}{\sqrt{1\times 3}+\sqrt{1\times\frac{1}{3}}}=\frac{0.5774}{1.7321+0.5774}\approx 0.25$$

为了识别随时间演变的网络论坛中的意见领袖，本文根据式(5)对每一时间窗口 T_k 内的静态图中的节点分别进行影响力计算，选取 T_k 内前 n 个影响力值最大的用户组成集合 U_{OL}^k，并对每一时间窗口内的结果进行匹配，即可跟踪随时间演变的意见领袖。具体算法如 Algorithm 1 所示。

Algorithm 1 Algorithm for Identification of opinion leaders

function SelectPotentialOpinionLeaders

1: for SF(T) do

2: for each node i over the graph G_k do

3: compute the Influence(i)^k

4: end for

5: sorting Influence(i)^k of all nodes N in a descending order

6: end if

7: return Top n potential opinion leader U_{OL}^k

8: end for

end function

function IdentifyOpinionLeaders

1: for each U_{OL}^k over the graph G_k do

2: $S=U_{OL}^0\cap U_{OL}^1\cap,\cdots,\cap U_{OL}^k$

3: end for

4: return each actual opinion leader $u_{ol}\in S$

end function

4 实验和结果分析

4.1 论坛数据获取

本实验以新浪网财经论坛(http://club.finance.sina.com.cn)的技术交流版块为研究对象，通过在配置为 Pentium (R) Dual-core CPU E5300 2.60 GHZ，1GHZ 内存的计算机上运行网络爬虫工具获取 2011 年 4 月－2011 年 10 月间的发帖数据。该时间区间共包含 5128 个帖子和 421 个参与发帖的用户。并按照以下方案建立论坛的网络图。

(1)如果发帖人对自己所发的帖子进行回复，则在本文中暂不建立节点的自我指向边；

(2)如果发帖人的帖子无回帖或只有自己回复，则删除该节点；

(3)如果回帖人 B 对发帖人 A 的帖子进行了回复，则认为 B 对 A 施加了一个影响。即在两个节点间建立由 B 指向 A 的边，边的权值根据回复的次数而定。

4.2 结果分析

为了验证上述算法的有效性，将采用上述算法获取到的结果与静态网络图(以连续 5 个月为时间窗口)中获取到的结果进行相似度值对比。

实验中，首先以连续 5 个月为时间窗口构建网络论坛的静态网络图(图 5 所示为利用 yEd Graph Editor 绘图工具绘制的静态网络图)，依据式(5)获取静态网络中的前 20 个意见领袖；然后依据 Algorithm1，以 4 月 1 日为起始点，以 15 天的时间周期为时间步长，构建出对应于 12 个不同时间窗口内的网络图。

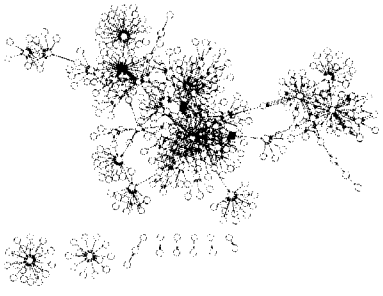

图 5 论坛静态网络图

为了选取算法中一个最优的 Top n 来提高识别精确度，实验分别取每一图中前 10 个和前 20 个潜在的意见领袖。将获取到的潜在意见领袖集合分别与图 5 中识别出的意见领袖集合进行 Jaccard 系数计算，建立对应于 Top10 和 Top20 的相似度值分布图，如图 6、图 7 所示。

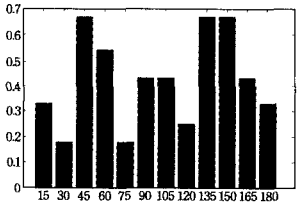

图 6 Top10 潜在意见领袖相似度值

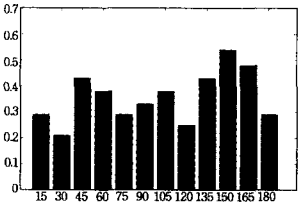

图 7 Top20 潜在意见领袖相似度值

通过观察图 6 和图 7 可以得出，两图的相似度值分布虽然比较相似，但图 6 相似度值平均要比图 7 高出 16%，表明采用前者的识别准确度更高。表 1 是结合式(5)和图(5)计算

出的影响力值排名前 10 的用户列表。

表 1　论坛静态网络图分析结果

UserId	Out-degree	Outgoing-clustering	Influence-value
历尽风雨见彩虹	59	0.98	44.50
财经小散	52	0.85	39.21
fcdsrgggggggg	50	0.78	37.70
frgtgt	48	0.74	36.19
云天梦	45	0.89	33.97
渐行渐远渐无言	41	0.75	30.94
俺 Q1395278391	40	0.87	30.22
飘飞的雪泥	38	0.72	28.68
2410897114hdd	37	0.72	27.93
伊凡童心	22	0.43	16.61

依据选用 Top $n=10$ 的算法对 12 个潜在意见领袖组成的集合相互进行匹配，获取的用户分别为：历尽风雨见彩虹，云天梦，俺 Q1395278391(它们在网络图中的位置用矩形块标识)。并将以上 12 个集合与表 1 中的结果进行 Jaccard 系数计算，将得到的相似度值建立如图 8 所示的相似度变化图。从图中可以发现，不仅各个时间步长上的相似度值低于 1，相邻步长的相似度值垂直变化也比较快。

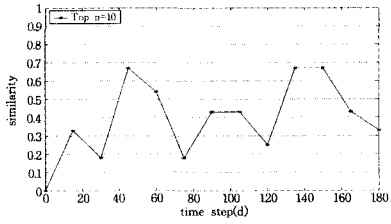

图 8　相似度变化图

以上实验结果说明，网络中用户的影响力随着时间的推移是动态变化的。同时也可以看出，由于噪声干扰的存在，相邻时间窗口内的识别结果存在偏差。但是，总体来看，相对于传统的基于静态网络图的论坛意见领袖发现方法，本文提出的算法不仅可以有效和准确地识别在某一短时间周期内变化的论坛意见领袖，而且可以将某一较长时间周期分割成不同的时间窗口，通过不同时间窗口内识别结果的相互匹配，较准确地发现该时间周期内的意见领袖。

结束语　在对网络论坛中意见领袖的识别研究中，本文提出了一个基于时间变化图的网络论坛意见领袖识别算法。实验表明，该算法与传统的基于静态网络图的识别方法相比，可以更准确地识别随时间的推移而变化的论坛意见领袖。

将来的工作主要包括：(1)将本文提出的算法中的网络论坛图的进化分割在数据的短期变化上，在降低噪声干扰的同时仍确保捕捉数据统计特性上的动向；(2)探讨结合多个时间窗口内的数据来计算在单个时间窗口的特征参数，使得所求特征参数随时间推移而平稳地变化，从而产生稳定的意见领袖识别结果。

参 考 文 献

[1] Hon Wai Lam, Chen Wu. Finding Influential eBay Buyers for Viral Marketing-A Conceptual Model of BuyerRank[C]// Proceedings of IEEE Conference on Commerce and Enterprise Computing. IEEE, 2009: 778-785

[2] Tang Xu-ning, Yang C C. Identifing influential users in an online healthcare social network[C]// Proc. IEEE Int. Conf. on Intelligence and Security Informatics, 2010 (ISI '10). May 2010: 43-48

[3] Bodendorf F, Kaiser C. Detecting Opinion Leaders and Trends in Online Communities[C]// 2010 Fourth International Conference on Digital Society. 2010: 124-129

[4] Rad A A, Benyoucef M. Towards Detecting Influential Users in Social Networks[C]// MCETECH 2011, LNBIP 78. 2011: 227-240

[5] Chen You, Cheng Xue-qi, Yang Sen. Finding High Quality Threads in Web Forums[J]. Journal of Software, 2011, 22(8): 1785-1804

[6] Casteigts A, Flocchini P, Quattrociocchi W, et al. Time-varying graphs and dynamic networks[R]. University of Carleton, 2010

[7] Esslimani I, Brun A, Boyer A. Detecting Leaders in Behavioral Networks[C]// 2010 International Conference on Advances in Social Networks Analysis and Mining. 2010: 281-285

[8] Zhou H. Scaling exponents and clustering coefficients of a growing random network[D]. Physical Review E 66. 2002

(上接第 32 页)

[4] Hou Meng-shu, Lu Xian-liang, Zhou Xu, et al. A trust model of P2P system based on confirmation theory [J]. Operating Systems Review, 2005, 39(1): 56-62

[5] Tian Chun-qi, Zou Shi-hong, Tian Hui-rong. A New Trust Model Based on Reputation and Risk Evaluation[J]. Journal of Electronics & Information Technology, 2007, 29(7): 1628-1632

[6] Li Xiong, Ling Liu. Peer Trust- supporting reputation-based trust for peer-to-peer election communities[J]. IEEE transactions on Knowledge and Data Engineering, 2004, 16(7): 843- 857

[7] Song S, Hwang K, Zhou R. Trusted P2P transactions with fuzzy reputation aggregation[J]. IEEE Internet Computing, 2005(6)

[8] Song Shan-shan, Huang Kai, Zhou Run-fang. Trusted P2P transactions with fuzzy reputation aggregation [J]. Internet Computing. 2005, 9(6): 24-34

[9] Altman J. PKI security for JXTA overlay networks[R]. TR-12-03-06. Palo Alto: Sun Microsystem, 2003

[10] 鲍翊平，姚莉，张维明，等. 对等网中基于种群进化的信誉模型[J]. 计算机科学，2011，38(1)：54-56

[11] Dou W, Wang H M, Jia Y, et al. A recommendation-based peer-to-peer trust model[J]. Journal of Software, 2004, 15(4): 571-583

[12] Zhang Q, Zhang X, Wen X Z, et al. Construction of peer-to-peer multiple-grain trust model[J]. Journal of Software, 2006, 17(1): 96-107

[13] Tian H, Zou S, Wang W. A Hierarchical reputation model for P2P networks[J]. Journal of Electronics and Information Technology, 2007(11)

[14] Swamynathan G, Zhao B Y, Almeroth K C. Decoupling service and feedback trust in a peer-to-peer reputation system[C]// Parallel and Distributed Processing and Applications Workshop 2005. Lecture Notes on Computer Science 3759, 2005: 82-90

[15] 封孝生，王桢文，黎湘运. P2P 中基于信任和属性的访问控制[J]. 计算机科学，2011，38(2)：28-31，41

[16] 黄骏虎，虞慧群. 一种基于信誉的 P2P 的评价模型[J]. 计算机科学，2011，38(Z10)：331-335

[17] 胡建理，吴泉源，周斌. P2P 环境下基于信誉的信任模型研究[J]. 计算机科学，2009，36(9)：1-6

[18] 杨超，刘念祖. P2P 环境下基于声誉的信任模型[J]. 计算机科学，2011，38(3)：131-135