

# 基于局部中心度的在线论坛意见领袖发现算法

俞 淮, 郑倩冰, 毛羽刚, 朱培栋

(国防科学技术大学 计算机学院, 湖南 长沙 410073)

**摘 要:**网络舆论形成过程中,其走向很大程度上受到意见领袖的影响。由于网络舆论的影响力不断增大,国内外学者也开始把研究重点放在网络论坛意见领袖上。从论坛帖子数据中提取回复关系,映射为发帖者和回帖者之间的关联关系,从而构造出一个社群网络。某个体的入度说明其被关注的程度,局部中心度直观地反映出与某个体直接联系的个体数目。基于入度和局部中心度的思想,并分析个体之间的交互行为,提出一种在线论坛的意见领袖发现算法。以某论坛为实验对象,找出其中的意见领袖,并通过分析实验结果验证文中算法的正确性。

**关键词:**意见领袖; 社会网络分析; 在线论坛; 局部中心度

**中图分类号:** TP393

**文献标识码:** A

**文章编号:** 1673-629X(2012)04-0009-03

## An Algorithm for Online Forum Opinion Leaders Discovery Based on Local Centrality

YU Huai, ZHENG Qian-bing, MAO Yu-gang, ZHU Pei-dong

(School of Computer Science, National University of Defense Technology, Changsha 410073, China)

**Abstract:** During the formation of network public opinion, opinion leaders in the network will greatly affect the direction of network public opinion. With the expansion of the influence of network public opinion, domestic and foreign scholars have begun to focus on the study of opinion leaders in online forums. Investigated an online forum and constructed a social network with the replying relations between posts and comments mapped to relations between posters and comment authors. Individual in-degree indicates the degree of concern while local centrality of an individual directly reflects the number of individuals that individual has direct contact with. Proposed an algorithm for discovering opinion leaders in online forums based on the idea of in-degree and local centrality, and analyse interactions between individuals. Took a forum as a subject, found out the opinion leaders, and verified the correctness of the algorithm by analyzing the experimental results.

**Key words:** opinion leader; social network analysis; online forum; local centrality

## 0 引 言

20世纪40年代,“意见领袖”的概念首先由美国哥伦比亚大学的传播学者保罗·拉扎斯费尔德提出<sup>[1]</sup>。在网络论坛中,只要能娴熟地操作网络、个性化地表达自己的观点、独到地参与话题讨论,任何网络用户都有可能成为意见领袖<sup>[2,3]</sup>。意见领袖的影响力<sup>[4,5]</sup>,贯穿整个网络舆论的形成过程,而且在其引导下,部分意见会演变成公众舆论<sup>[6]</sup>。因此,研究论坛的意见领袖具有重大的意义,也引起了国内外学者的广泛关注。

随着 Web,电子商务应用和在线社交网络的高度发展,了解信息流及其影响变得至关重要。事实上,

Web 2.0 的成功,带动了比如在线论坛、社交网站、博客、wiki 等虚拟社区的出现。对这些虚拟社区进行社会网络分析<sup>[7,8]</sup>,一个重要的方面是发现其中的领导者。然后通过分析个体之间的相互影响,达到研究社会网络属性,并预测其演化的目的<sup>[9]</sup>。

Amit Goyal<sup>[3]</sup>等从病毒式营销中受到启发,设计基于频繁模式挖掘的算法发现社会网络的领袖和部落。其社群网络的构建是基于用户在社交网站中的行为,而某个用户的影响力网络是基于其行为传播扩散的时间,判断用户是否为领袖的依据是该用户的影响力网络的大小。高俊波<sup>[10]</sup>等利用图论中的网络平均路径长度算法,比较去除每个节点前后网络的平均路径长度差,发现了网络论坛的意见领袖。但是他们设计的算法只分析最大连通图,却忽视了对其他连通分支的分析,将会导致结果的不准确。王钰<sup>[11]</sup>等设计基于 EM 算法的用户聚类算法,从实际的帖子数据中提取向量数据集,并基于聚类结果筛选出最符合意见领袖

收稿日期:2011-09-02;修回日期:2011-12-06

基金项目:国家自然科学基金项目(60873214,61170285)

作者简介:俞 淮(1984-),男,江西婺源人,硕士,主要研究领域为社会网络分析。

群体的子类。但是其用平均被回复长度与回复长度之差作为筛选意见领袖的条件准确度不高。

文中提出的算法是从论坛帖子数据中提取回复关系,映射为发帖者和回帖者之间的关联关系,从而构造出一个社群网络。基于入度和局部中心度筛选出可能的意见领袖集,然后分析可能的意见领袖与其他个体的交互行为,以被回复长度和交互的流量为标准,找出真正的意见领袖。该算法既考虑社群网络的结构特性,又结合了个体之间的交互行为,找出的意见领袖更符合实际。实验中,以中华网论坛为例,运用该方法提取出意见领袖,并通过分析实验结果验证文中算法的正确性。

## 1 基于局部中心度的在线论坛意见领袖发现算法

该算法的输入是在线论坛用户所构成的社群网络  $G=(N,E)$ , 所以,首先从论坛的实际帖子数据中,提取用户之间的交互关系,构建社群网络。然后根据定义的意见领袖选取标准,设计算法的具体过程分为两步:

第一步,根据节点的入度和局部中心度,筛选出可能的意见领袖集;

第二步,根据被回复长度和流量从中找出真正的意见领袖。

### 1.1 社群网络的构建

根据在线论坛的特点,对构建的社群网络  $G=(N,E)$  定义如下:

- (1) 节点:网络论坛中的用户;
- (2) 边:若两个节点之间有回复关系就加一条边,由回帖者指向发帖者;
- (3) 边的权值:某个节点的被回复长度与回复长度之和,文中称为该节点的流量。

这种由人及其互动关系所组成的网络是一个复杂网络,文中定义该复杂网络图是一个有向加权图。

### 1.2 意见领袖的选取

Everett M. Rogers<sup>[12]</sup>把意见领袖的特征总结为广泛的知识面,一定的言论引导作用,丰富的社会资源,并且有数目不少的“粉丝”。同样,在网络论坛中,处于领袖地位的核心成员必定视野开阔,积极活跃,并且观点个性鲜明能引起广泛关注。因此,他们在论坛中的活跃程度及发帖的影响力都是超群的。

基于上文所构建的社群网络,可以得到一个以论坛作者为节点、他们之间的回复关系为边、两节点之间流量为权值的网络。意见领袖的获取过程就是在网络中找寻关键节点的过程。意见领袖的观点是能够引起广泛关注的,所以该节点的入度一定要大。局部中

心度能反映出一个节点在网络中所处的重要程度,因此可能的意见领袖的局部中心度也一定要大,当两个节点入度相同时,局部中心度大的排名在前。另外,在流量方面,可能的意见领袖的被回复的长度应该比较大,当两个节点被回复的长度相同时,流量大的排名在前。

### 1.3 算法描述

局部中心度:某点的“局部中心度”通常用该点的“度数”就可以直观地描述,也就是与这个个体直接联系的个体数目。可以想象,如果某个个体的度数很大,往往说明该个体处于某种中心的位置上,与其他很多个体产生关系。

如算法 1 所示,基于局部中心度的思想设计本算法,具体过程分两步:

1) 找出可能的意见领袖集。

计算图  $G=(N,E)$  中各个节点  $u_a$  的入度  $D_i(u_a)$  和局部中心度  $D_{lc}(u_a)$ , 按照入度的大小降序排列,如果两个节点的入度相同则按局部中心度降序排列,取 TopM 个节点作为可能的意见领袖集  $U_{PL}$ 。

2) 发现意见领袖。

对可能的意见领袖集中的每个节点  $u_{pl} \in U_{PL}$ , 计算其被回复长度  $Lr(u_{pl})$  和流量  $Fl(u_{pl})$ 。然后按照被回复长度进行降序排列,如果两个节点的被回复长度相同,则按流量降序排列,即可得到 TopN 个真正的意见领袖集  $U_L$ 。

算法 1. 基于局部中心度的在线论坛意见领袖发现算法。

```

Input:  $G=(N,E)$ 
Output:  $U_L$ , which contains TopN opinion leaders
1.  $U_{PL} \leftarrow \emptyset, U_L \leftarrow \emptyset$ 
2. for each node  $u_a$  over the graph  $G$  do
3. Evaluate “In-degree”  $D_i(u_a)$  and “Local centrality degree”  $D_{lc}(u_a)$ 
4. end for
5. Sorting In-degree  $D_i$  of all nodes  $N$  in a descending order
6. for each node  $u_i$  and  $u_j$  in  $N$  do
7. if  $D_i(u_i) = D_i(u_j)$  then
8. Sorting them with  $D_{lc}(u_i)$  and  $D_{lc}(u_j)$  in a descending order
9. end if
10. end for
11.  $U_{PL} \leftarrow N(\text{TopM})$ 
12. for each potential opinion leader  $u_{pl} \in U_{PL}$  do
13. Evaluate “Length was replied”  $Lr(u_{pl})$  and “Flow”  $Fl(u_{pl})$ 

```

```
14. end for
15. Sorting Lr of all nodes in  $U_{PL}$  in a descending order
16. for each node  $u_i$  and  $u_j$  in  $U_{PL}$  do
17.   if  $Lr(u_i) = Lr(u_j)$  then
18.     Sorting them with  $Fl(u_i)$  and  $Fl(u_j)$  in a desending order
19.   end if
20. end for
21.  $U_L \leftarrow U_{PL}(\text{TopN})$ 
22. return  $U_L$ 
```

2 实 验

2.1 现实数据集

文中以中华网论坛为实验对象,选取其中的中华军备板块 2011 7 月 7 日的所有帖子作为数据源,包括了 96 个帖子,680 位论坛用户。然后提取每个帖子中的帖子时间、发帖人、帖子内容长度、回帖人及回帖内容长度组成数据集。

2.2 结果分析

根据输入数据集构建社群网络图  $G = (N, E)$ , 计算各个节点的入度和局部中心度,按照入度的大小降序排列,如果两个节点的入度相同则按局部中心度降序排列,得到 Top10 的可能意见领袖集  $U_{PL}$  如表 1:

表 1 Top10 可能的意见领袖集  $U_{PL}$

序号	用户名	入度	局部中心度
1	苍鹰的翅膀	638	638
2	草思念	250	250
3	噱头	199	199
4	笑谦	79	79
5	醉酒豪情	75	78
6	莫回刀	71	71
7	xukon	65	65
8	航母五星上将	54	60
9	读破万卷	51	56
10	照明 168	51	52

在上表中可以看出,“读破万卷”和“照明 168”的入度相同,但是“读破万卷”的局部中心度为 56,即回复了别人的帖子 5 次,说明在论坛的活跃度相对较高。因此,按照局部中心度的降序排列,“读破万卷”的排名在前。

对可能的意见领袖集  $U_{PL}$  中的每个节点计算被回复长度和流量,然后按照被回复长度进行降序排列,如果两个节点的被回复长度相同,则按流量降序排列,得到 Top5 真正的论坛意见领袖集  $U_L$  如表 2:

中心度在一定程度上能反映出某个节点在网络中所处的中心位置,从结果可以看出真正的意见领袖集

与可能的意见领袖集的前五名基本相同,只是“醉酒豪情”和“笑谦”的排名有变化。分析论坛的数据得知,“笑谦”虽然局部中心度较大,但是只发了 1 帖,而且回复的长度较短,而“醉酒豪情”共发了 8 帖,反映出其在论坛上比较活跃,而且其中 3 帖关注度较高,回复的长度较长,更符合意见领袖的特征,从而也验证了算法的准确性。

表 2 Top5 真正的意见领袖集  $U_L$

序号	用户名	回复长度	被回复长度	流量
1	苍鹰的翅膀	0	10847	10847
2	草思念	0	3759	3759
3	噱头	0	2783	2783
4	醉酒豪情	10	977	987
5	笑谦	0	542	542

在线论坛中用户的个人资料也能反映出用户在论坛上的活跃程度,表 3 是真正的意见领袖集  $U_L$  中成员的个人资料:

表 3 真正的意见领袖集  $U_L$  成员的个人资料

序号	用户名	积分	等级	级别
1	苍鹰的翅膀	55013 分	16	少将
2	草思念	36589 分	15	大校
3	噱头	185201 分	19	元帅
4	醉酒豪情	88805 分	17	中将
5	笑谦	445264 分	20	大元帅

从表 3 中可以看出,这 5 人的等级都较高,积分也很高,是论坛的核心人物,这也从侧面验证了算法的准确性。至于排名的顺序,这是因为数据集的帖子数较少,而用户的个人等级和积分是个积累的过程。要提高结果的准确度,可以增加数据集的大小,分析的方法是一样的。

3 结束语

文中提出了基于局部中心度的在线论坛意见领袖发现算法,首先从网络结构本身,以入度和局部中心度为标准筛选可能的意见领袖集,然后从节点之间的交互,以被回复的长度和流量为依据找出真正的意见领袖。并以中华网论坛的中华军备板块为实验对象,提取用户之间的回复关系,构建社群网络,采用文中的意见领袖发现算法,找出真正的意见领袖。然后对找出的意见领袖进行分析,验证算法的正确性。

分析和研究论坛意见领袖是对单个关系网络进行社会网络分析的一个典型应用,下一步的工作是对综合分析多个关系网络的相关技术进行进一步的研究。

参考文献:

[1] Ohsawa Y. Chance discoveries for making decisions in complex real world [J]. New Generation Computing, 2002, 20

行向量重构出的图像数据值  $\hat{X}_h$ ;

步骤4: 均衡步骤2和步骤3中的图像数据,得到

$$\text{最终重构结果: } \hat{X} = \frac{(\hat{X}_h + \hat{X}_v)}{2}.$$

本设计中选取图像为 Lena(320 \* 240), 压缩率为 1/2, 采用 OMP 重构算法对图像进行重构。图5给出了未经图像均衡与图像经过均衡后的重构图像对比。



(a) 未经图像均衡行列值

(b) 图像均衡行列值

图5 未经图像均衡与图像经过均衡后的重构图像

从图5(a)可见, 未经图像均衡行列值重构的图像有部分大块的白点, 而图5(b)为经过图像均衡行列值的重构图像, 没有出现这些白色的块点, 经过图像均衡行列值的算法优于未经图像均衡行列值的传统算法。

#### 4 结束语

压缩感知理论作为一个新兴的信号处理理论在图像处理中有广阔的应用前景。文中通过介绍压缩感知的基本理论, 利用以 NIOS II 软核为核心组成的数字信号处理平台, 将图像采集、压缩与重构有效地整合为一体。针对 OMP 重构算法在实际工程中计算时需要大量存储空间并耗时巨大的问题, 文中给出了图像分块压缩的改进方案; 针对 OMP 算法重构时图像列与列之间数据相关性被割裂的现象, 给出了具体的图像均衡行列值算法。实际系统运行结果显示两种改进方案均

取得良好效果, 基于 NIOS II 的图像压缩感知系统适合于实时性要求高、数据信息量大的终端系统。

#### 参考文献:

- [1] Donoho D. Compressed Sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.
- [2] Candes E, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information[J]. IEEE Transactions on Information Theory, 2006, 52(2): 489-509.
- [3] 石光明, 刘丹华, 高大化, 等. 压缩感知理论及其研究进展[J]. 电子学报, 2009, 37(5): 1070-1081.
- [4] 戴琼海, 付长军, 季向阳. 压缩感知理论[J]. 计算机学报, 2011, 34(3): 425-434.
- [5] 许芳, 席毅, 陈虹, 等. 基于 FPGA/Nios-II 的矩阵运算硬件加速器设计[J]. 电子测量与仪器学报, 2011, 25(4): 377-383.
- [6] Baraniuk R. A lecture on compressive sensing[J]. IEEE Signal Processing Magazine, 2007, 24(4): 118-121.
- [7] Chen S B, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit[J]. SIAM Journal on Scientific Computing, 1998, 20(1): 33-61.
- [8] Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit[J]. IEEE Transactions on Information Theory, 2007, 53(12): 4655-4666.
- [9] Antonini M, Barlaud M, Mathieu P, et al. Image coding using wavelet transform[J]. IEEE Transactions on Image Processing, 1992, 1(2): 205-220.
- [10] Carvajalino D, Sapiro G. Learning to Sense Sparse Signals; Simultaneous Sensing Matrix and Sparsifying Dictionary Optimization[J]. IEEE Transactions on Image Processing, 2009, 18(7): 1395-1408.

(上接第11页)

(2): 143-164.

- [2] Esslimani I, Brun A, Boyer A. From social networks to behavioral networks in recommender systems[C]//Proceedings of the 2009 International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Washington, DC, USA; IEEE Computer Society, 2009: 143-148.
- [3] Goyal A, Bonchi F, Lakshmanan L V. Discovering leaders from community actions[C]//Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08). New York, USA; ACM, 2008: 499-508.
- [4] 王陆, 马如霞. 意见领袖在虚拟学习社区社会网络中的作用[J]. 电化教育研究, 2009(1): 54-58.
- [5] 王丽. 虚拟社群中意见领袖的传播角色[J]. 新闻界, 2006(3): 50-51.
- [6] 胡勇, 张翀斌. 网络舆论形成过程中意见领袖的形成模型研究[J]. 四川大学学报(自然科学版), 2008, 45(2):

347-351.

- [7] 刘军. 社会网络分析导论[M]. 北京: 社会科学文献出版社, 2004.
- [8] 彭小川, 毛晓丹. BBS 群体特征的社会网络分析[J]. 青年研究, 2004(4): 39-44.
- [9] Esslimani I, Brun A, Boyer A. Detecting Leaders in Behavioral Networks[C]//Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Washington, DC, USA; IEEE Computer Society, 2010: 282-285.
- [10] 高俊波, 杨静. 在线论坛中的意见领袖分析[J]. 电子科技大学学报, 2007, 36(6): 1249-1252.
- [11] 王钰, 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): 44-46.
- [12] Rogers E M. Diffusion of innovations[M]. 4th ed. New York, USA; the Free Press, 1995.