文章编号：1003-0077(2018)02-0129-10

基于多特征信息传播模型的微博意见领袖挖掘

张 米[1]，张 晖[2]，杨春明[1]，李 波[1,3]，赵旭剑[1]

(1. 西南科技大学 计算机科学与技术学院，四川 绵阳 621010；
2. 西南科技大学 理学院，四川 绵阳 621010；
3. 中国科学技术大学 计算机科学与技术学院，安徽 合肥 230027)

摘 要：在线社交网络中的意见领袖通常是指在社交网络的信息传播中具有较大社会影响力的个体。针对当前意见领袖挖掘方法中只考虑社交网络的拓扑结构和节点的个体属性，缺乏信息传播中交互特征的问题，该文提出了基于扩展独立级联模型，并融入网络结构特征、个体属性和行为特征的意见领袖挖掘模型(extended independent cascade, EIC)。该模型以个体属性、个体在信息传播过程中的交互行为建立加权的传播网络，利用改进的 CELF (cost effective lazy forward)算法，挖掘网络中影响力较大的个体。通过实验验证，在意见领袖的扩展核心率指标上，该算法优于拓扑结构类算法，且具有较好的稳定性，同时并未降低意见领袖的传播范围。

关键词：独立级联模型；信息传播；传播模型；意见领袖

中图法分类号：TP391 **文献标识码**：A

Microblog Opinion Leader Mining Based on a Multi-feature Information Diffusion Model

ZHANG Mi[1], ZHANG Hui[2], YANG Chunming[1], LI Bo[1,3], ZHAO Xujian[1]

(1. School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China;
2. School of Science, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China;
3. School of Computer and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China)

Abstract: Opinion leaders in online social network are those who have great social influence. Current opinion leader mining methods consider only the topological structure of a social network and the node's attributes, neglecting the interaction in information diffusion. This paper proposes an opinion leader mining model based on the independent cascade model named EIC (Extended Independent Cascade), which incorporates the network structure, the node attributes, and the user behavior characteristics to build a weighted diffusion network. Experimental results of real data collected from Sina Weibo show that the proposed algorithm is superior to the topological structure algorithms in the extended core rate of opinion leaders, without under-estimate the the scope of influence.

Key words: independent cascade model; information diffusion; diffusion model; opinion leader

0 引言

社交网络是通过社会成员之间的交互形成的相对稳定的社会结构，具有复杂的网络结构和动态传播机制，对信息的传播和扩散起着至关重要的作用。在某一事件的传播过程中，社交网络中对事件的传播与发展具有关键导向作用且社会影响力较大的部分节点，被称作“意见领袖”(opinion leader)。越来越多的研究表明，意见领袖在信息传播扩散、网络舆情监控、网络口碑效应等社会现象中具有不可估量的作用[1-2]。因此，社交网络中意见领袖的识别与挖掘逐渐成为当前的研究热点。

目前，针对意见领袖挖掘的研究多集中在社交

收稿日期：2017-06-06 定稿日期：2017-07-20

基金项目：四川图书情报研究中心资助课题(SCTQ2016YB13)；四川省军民融合研究院开放基金(18sxb028，18sxb017)

网络的拓扑结构特征上，重点关注节点的重要性。然而微博作为当今社交网络中最盛行的网络交流平台之一，与其他社交应用相比，其用户属性不同，信息交互方式也不同，并且意见领袖所产生的社会影响力与特定的话题紧密相关，不同话题下表现的影响力也大相径庭。在某个特定的话题下，一个拥有大量粉丝的用户，若其活跃度很低，也很难在该话题下拥有与粉丝数正相关的社会影响力。基于Twitter的研究也表明：在信息的传播过程中，用户的影响力与其粉丝数量成弱相关关系[3-4]。因此，在挖掘意见领袖时，还需要强调微博信息的话题特征、网络节点的属性特征、行为特征以及传播规律，而非仅仅注重网络的拓扑结构。

针对上述问题，本文提出了一个融合网络结构特征、用户个体属性以及用户交互行为的多特征信息传播模型EIC，在信息动态传播的过程中定量分析用户间的影响能力，以此挖掘微博特定话题下最具影响力的节点集合。

1 相关工作

目前社交网络中意见领袖的挖掘重点关注网络拓扑结构、用户内容、用户交互行为等多个方面，主要分为三类方法。

（1）用户属性分析法

基于对用户节点自身属性的分析。许丹青等[5]从用户的阅读概率角度引入用户的发文行为、浏览行为与标签社区小世界属性等对用户的影响力进行建模。Lin等[6]通过对用户多种属性的分析，提出了挖掘Facebook粉丝页面意见领袖的聚类算法。用户属性分析法在构建属性指标时，相应指标的权重因素需要事先给出，一定程度上取决于研究者的主观意向，因而挖掘结果波动程度较大。此外，该类方法普遍缺乏对网络拓扑结构的考虑，存在一定的局限性。

（2）网络结构分析法

侧重从网络结构的角度挖掘意见领袖。Weng等[7]分析相关话题下用户间的链接关系，提出了TwitterRank算法来挖掘特定话题下具有影响力的Twitter用户。吴岘辉等[8]考虑用户固有属性和网络结构，提出了OpinionLeaderRank算法来挖掘微博话题的意见领袖。网络结构分析法着重考虑网络拓扑结构，对于用户个体属性以及行为特征考虑不够全面。

（3）信息交互分析法

通过分析用户发表信息的影响力及传播特性来反映用户的影响力。朱玉婷等[9]考虑节点间的社交关系以及主题分布，构建了基于主题的信息传播模型TPM。Yang等[10]考虑非相邻节点情况以及信息传播延迟，提出了基于定量方法的中间关系算法来识别意见领袖节点。Jendoubi等[11]融合用户在网络结构中的重要性以及用户消息的普及范围，提出了基于Twitter社交网络的证据影响最大化模型。信息交互分析法重点关注传播模型在特定网络结构上的扩散，较少考虑节点属性和行为特征。

在采用信息交互分析法挖掘意见领袖时，社交网络中节点的影响力表现在节点行为在特定模型下的传播范围，而节点行为的传播主要依赖影响传播概率。在传统独立级联模型中[12]，影响传播概率通常由固定常数表示，或由网络拓扑结构特征得出。但在社交网络中，节点的影响力不仅与节点的个体差异有关，如用户的基本特征、行为特征等，还与节点之间的关系特征有关。此外，由于微博具有显著的话题特征，导致意见领袖具有话题依赖性，因此，在挖掘微博意见领袖时，话题特征是不可忽略的因素。

2 基于传播模型的意见领袖挖掘

2.1 微博有向加权图的构建

微博话题信息传播的过程可看作一幅有向加权网络图$G(V, E, P, W)$，如图1所示。图中，圆圈表示参与话题消息发表、转发、评论的微博用户，构成了用户节点集合V；有向线段表示信息在社交网络下用户交互过程中的传播方向，构成了有向边集合E。w_i、w_j等表示相应下标节点i、j的权重，构成了节点权重集合P；w_{ij}、w_{jk}等表示相应下标有向边e_{ij}、e_{jk}的权重，构成了有向边权重集合W。

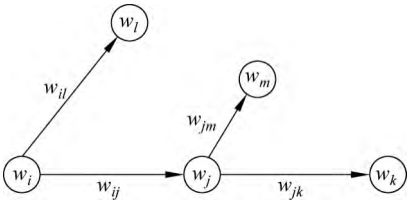

图1 微博信息传播有向加权网络图G

2.2　多特征信息传播模型

在独立级联模型中[12]，用户间的影响传播概率通常在一定范围内随机取值，忽略了社交网络中影响传播概率的差异性。而在真实的信息传播过程中，节点间的影响传播概率与节点特征、节点间交互特征有关。因此，本文将信息传播者 i 的节点特征 w_i、信息接受者 j 的节点特征 w_j 以及 i、j 间的交互特征 w_{ij} 融合成信息传播模型中的影响传播概率，如图 2 所示。

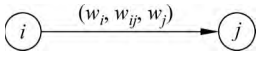

图 2　影响传播概率的权重向量

2.2.1　用户节点特征提取

根据意见领袖的定义，本文将从以下三个角度评估微博网络中的用户节点：

(1) 用户影响力(influence)：用户发表的微博消息对其他用户造成的影响。在微博平台下，节点 i 的影响力可表示为式(1)：

$$\mathrm{influ}(i) = \lambda_1 N_{\mathrm{weiborep}} + (1-\lambda_1) N_{\mathrm{weibocom}} \tag{1}$$

式(1)中，N_{weiborep} 表示用户发表的信息被转发的次数，用来衡量用户发布的微博的价值大小；N_{weibocom} 表示用户发表的信息被评论的次数，用来衡量用户发布的微博的交流与传播的热度；λ_1 表示信息被转发数属性的比例系数。

(2) 用户活跃度(activity)：体现用户节点在话题传播过程中主动参与的积极性。在微博平台下，节点 i 的活跃度可表示为式(2)：

$$\begin{aligned}\mathrm{activ}(i) =& \lambda_2 N_{\mathrm{createweibo}} + \lambda_3 N_{\mathrm{repweibo}} \\ &+ (1-\lambda_2-\lambda_3) N_{\mathrm{comweibo}}\end{aligned} \tag{2}$$

式(2)中，$N_{\mathrm{createweibo}}$ 表示原创微博数量，用来衡量用户表达思想、发表观点的意愿；N_{repweibo} 表示转发微博数，用来衡量用户传播、扩散消息的积极性；N_{comweibo} 表示评论微博数，用来衡量用户对其他用户的言论关注度；λ_2、λ_3 表示相应属性的比例系数。

(3) 用户权威度(authority)：衡量用户的知名度以及影响力。在微博平台下，节点 i 的权威度可表示为式(3)：

$$\mathrm{auth}(i) = \lambda_4 N_{\mathrm{fans}} + \lambda_5 N_{\mathrm{mutual}} + (1-\lambda_4-\lambda_5) N_{\mathrm{attention}} \tag{3}$$

式(3)中，N_{fans} 表示用户的粉丝数，用来衡量用户受关注的程度，以及受其影响的潜在用户的数目；N_{mutual} 表示相互关注数，用来衡量用户稳定的信息来源和扩散途径的多少；$N_{\mathrm{attention}}$ 表示关注数，用来衡量用户信息传播来源的广泛性以及可靠性；λ_4、λ_5 表示相应属性的比例系数。

节点的详细属性衡量指标如表 1 所示。

表 1　多特征信息传播模型衡量指标

	一级属性	二级属性
用户属性特征	影响力	被转发数
		被评论数
	活跃度	原创微博数
		转发数
		评论数
	权威度	粉丝数
		关注数
		互关数
用户交互特征	转发	—
	评论	—

从影响力 $\mathrm{influ}(i)$、活跃度 $\mathrm{activ}(i)$ 和权威度 $\mathrm{auth}(i)$ 三个角度评估节点 i 后，节点权重 w_i 可表示为式(4)：

$$\begin{aligned}w_i =& \mu_1 \mathrm{influ}(i) + \mu_2 \mathrm{activ}(i) \\ &+ (1-\mu_1-\mu_2) \mathrm{auth}(i)\end{aligned} \tag{4}$$

式(4)中，μ_1、μ_2 表示相应属性的比例系数。

2.2.2　边的特征提取

本文根据用户间的转发、评论关系构建微博网络图中的有向边，边权重代表联系紧密强度。因此，在计算边权重时，考虑了节点间的转发、评论次数，次数越多，节点间的联系越密切，越容易受邻居节点影响。在微博平台中，节点 i 与节点 j 间的有向边 e_{ij} 的权重 w_{ij} 可表示为式(5)：

$$w_{ij} = \eta N_{\mathrm{rep}} + (1-\eta) N_{\mathrm{com}} \tag{5}$$

式(5)中，N_{rep} 表示节点 j 转发节点 i 的微博数；N_{com} 表示节点 j 评论节点 i 的微博数；η 表示相应属性的比例系数。

2.2.3　影响传播概率的计算

(1) 指标归一化

影响传播概率由数据的相关属性以及各属性的权重系数决定。相关属性的度量是影响精确度的主要因素。在实际数据中，各属性取值范围不一，需要进一步做标准化处理，本文采用 Min-Max 标准化方法，将不同属性的取值映射到[0，1]之间。

(2) AHP确定权重

各属性的权重系数是影响精确度的另一因素，本文采用定性与定量相结合的AHP多准则决策方法对其进行确定[12]。

首先根据“一样重要”“稍微重要”“明显重要”“重要得多”“极端重要”等标准对各层次内属性指标进行两两比较，采用1-9标度法，确定每个层次内各指标的相对重要性，得到判定矩阵，如表2所示；然后进行一致性检验；再采用方根法计算判定矩阵的特征向量，得到各层次内属性指标的权重系数，进而分别计算出影响力、活跃度、权威度、用户节点权重、边权重以及影响传播概率六个层次上的权重系数，结果如表3所示。

表2 节点一级属性判定矩阵

属性指标	影响力	活跃度	权威度
影响力	1	2	4
活跃度	1/2	1	3
权威度	1/4	1/3	1

表3 影响传播概率评价指标与权重

总指标	一级指标	权重	二级指标	权重
用户节点权重	影响力	0.449 7	被转发数	0.615 5
			被评论数	0.384 5
	活跃度	0.391 2	原创数	0.552 2
			转发数	0.285 7
			评论数	0.162 1
	权威度	0.159 1	粉丝数	0.400 9
			互关数	0.347 9
			关注数	0.251 2
边权重	转发交互	0.487 2		
	评论交互	0.512 8		
影响传播概率	传播者	0.189 3		
	接受者	0.189 3		
	交互	0.621 4		

(3) 传播概率计算

在信息的传播过程中，分析信息传播者和信息接受者的结构特征、个体特征以及行为特征，根据微博数据计算出信息传播者的节点权重 w_i、信息接受者的节点权重 w_j、以及两者间边权重 w_{ij}，节点间一对一的影响传播概率可以表示成以上三个要素的线性组合。具体计算公式可表示为式(6)：

$$p_{ij} = \alpha_1 w_i + \alpha_2 w_j + (1-\alpha_1-\alpha_2) w_{ij} \quad (6)$$

式(10)中，α_1、α_2表示相应属性的比例系数。

2.3 基于EIC模型的意见领袖挖掘

2.3.1 意见领袖的定义

基于EIC模型的意见领袖挖掘是在给定的EIC模型下，获取一个大小为 k 的节点集合，使得该集合在微博有向加权图 $G(V, E, P, W)$ 中可以将信息传播扩散到最多的节点，这 k 个节点即是所要挖掘的意见领袖。

2.3.2 意见领袖的挖掘

Kempe、Kleinberg和Tardos证明在多种传播模型下，挖掘影响力最大的种子节点是NP-hard问题[13]。因此，在衡量种子选择算法的近似准确性以及时间复杂度后，本文借鉴Leskovec等提出的CELF算法[14]，并进行优化，如算法1所示。在计算节点的边际收益时，采用了EIC模型中融合用户多种属性的动态性影响传播概率代替以往传播模型中的静态传播概率。差异性的传播概率同时考虑了信息传播者传播信息的能力和信息接受者接受信息的意愿，更加接近真实社交网络中的影响力。

算法1 基于EIC模型的CELF算法

输入：微博有向加权图 $G(V, E, P, W)$，话题特征 T，意见领袖个数 k
输出：话题相关的意见领袖集合 S
算法描述：

```
1 begin
2   S=ϕ;
3   //S: the set of seed nodes
4   Q=ϕ;
5   //Q: sorted list in decreasing order according to the
      marginal gain of nodes
6   for each v∈ V do
7     marginalGain(v);
8     //marginalGain()estimate the marginal gain of v
9     Q.add(v);
10  end
11  nodeMax←Q.pop();
12  S.add(nodeMax);
13  while |S|≤k do
14    nodeMax←Q.pop();
15    updateMarginalGain(nodeMax);
16    if nodeMax.MG≥Q.getFirst().MG then
        S.add(nodeMax);
17    else Q.add(nodeMax);
18  end
19 end
```

算法第 6～10 行：迭代计算每个节点的边际收益，并根据边际效益值的降序排列将节点保存在一个有序列表中；第 11～12 行：选取有序列表的 top 节点作为首个种子节点；第 13～18 行：重新评估列表 top 节点的边际效益，并对列表重新排序，如果 top 节点依旧保持在第一个位置，则该节点为种子节点，否则重新评估新列表的 top 节点的边际效益，直到获取到 k 个种子节点为止。

3　实验

3.1　实验数据

本文基于新浪微博的开放 API 获取“嫦娥三号”“两会房价”“两会雾霾”及“单独二胎”四个代表性话题在一定时间范围内的相关传播数据，挖掘特定话题下的意见领袖集合。其中每个话题均包含用户属性信息、用户交互信息以及微博信息，详细信息见表 4。

表 4　实验数据集信息

话题	原创微博数	转发微博数	评论数	用户数	时间范围
嫦娥三号	2 696	2 251	1 138	4 933	2013.12.02-2013.12.08
两会房价	13 111	35 880	35 105	67 646	2014.03.03-2014.04.19
两会雾霾	6 997	7 731	7 527	17 151	2014.03.03-2014.04.22
单独二胎	16 676	34 045	131 699	125 584	2012.08.13-2013.11.21

为了验证本文模型的普适性，对上述数据构建的微博有向加权图进行特征统计，统计结果如表 5 所示。

表 5　实验网络图特征统计

话题	平均度	平均聚类系数	网络直径	平均路径长度
嫦娥三号	2.067	0.012	3	1.308
两会房价	2.035	0.007	12	5.014
两会雾霾	2.034	0.033	9	1.657
单独二胎	2.26	0.021	10	2.526

其中，平均度是网络中所有节点度的平均值，而节点的度是与该节点连接的边数；平均聚类系数是网络中所有节点聚类系数的平均值，而节点的聚类系数是节点与其邻居节点之间存在的边数与最多可能有的边数之比；这两个特征衡量了网络的连通性。网络直径是网络中任意两个节点之间距离的最大值；平均路径长度是网络中所有节点对之间距离的平均值；这两个特征衡量了网络的传输效率。从表 5 可以看出，四个网络的各个统计特征均有所差异，造成各个网络的紧密程度也大小不一。

3.2　对比算法

为验证本文提出的 EIC 算法的性能，将它与多种拓扑结构类算法进行比较，分析网络拓扑结构特征、行为特征和用户属性特征对意见领袖挖掘结果的影响。实验对比的算法及参数设置描述如下：

（1）PR 算法（PageRank）：根据网页之间的超链接关系，利用均分“投票”思想标识网页的等级或重要性，α 值设置为 0.85。

（2）OLR 算法（OpinionLeaderRank）[8]：根据用户间话题相关的信息交互关系构建用户行为网络图，利用随机游走思想挖掘微博中意见领袖，α 值设置为 0.55。

（3）MR 算法（Microblog-Rank）[15]：依据用户间的评论关系构建网络图，扩展 PageRank 算法来分析用户的网络重要性，α 值设置为 0.55。

3.3　评价指标

意见领袖的评估尚未有公认、统一的标准体系。本文根据意见领袖的定义，采用扩展核心率和影响传播范围两种评估指标，从个体属性、网络结构以及信息传播扩散能力多个方面对意见领袖进行综合评价[8,12]。既考虑了意见领袖的静态特征，又衡量了意见领袖的动态传播能力。同时，采用意见领袖一致性指标评估意见领袖的话题依赖程度[16]。

（1）扩展核心率 ECR（extended core ratio）

从用户的个体属性和网络结构两个方面衡量意见领袖的代表性，计算公式如下：

$$\mathrm{ECR}_i = \varepsilon a_i + (1-\varepsilon) b_i \quad i \in V \tag{7}$$

$$a_i = \frac{w_i}{\sum_{j=1}^{N} w_j} \quad i \in V \tag{8}$$

$$b_i = \frac{\varphi \sum_{e_{ji} \in E} w_j w_{ji} + (1-\varphi) \sum_{e_{ij} \in E} w_j w_{ji}}{\sum_{e_{jk} \in E} [\varphi w_j w_{jk} + (1-\varphi) w_k w_{jk}]} \tag{9}$$

式中，a_i 表示用户属性权重比率；b_i 表示网络核心率；ε 表示用户属性权重比率在扩展核心率中占的比重；φ 表示节点的入边在网络核心率中所占的比重。

(2) 影响传播范围 IS(influence spread)

从信息动态扩散的角度衡量意见领袖的感染力。基于微博社交网络的特点,将微博网络中用户 i 的影响传播范围定义为:在一定时间范围内,转发或评论用户 i 微博的用户数。用户 i 的影响传播范围可表示为式(10):

$$\mathrm{IS}(i)=\sum_{j\in V}\mathrm{Rep}(i,j)+\sum_{k\in V}\mathrm{Com}(i,k)\quad i\in V \tag{10}$$

式(10)中,$\mathrm{Rep}(i,j)$表示用户 i 的微博被用户 j 转发;$\mathrm{Com}(i,k)$表示用户 i 的微博被用户 k 评论。为获得用户较为精确的影响传播范围,本文采用 Monte-Carlo 方法模拟 10 000 次传播过程,取 10 000 次平均值作为用户最终的影响传播范围。

(3) 意见领袖一致性 LC(leader consistent)

衡量不同话题下意见领袖集合的交叉程度,计算公式如下:

$$\mathrm{LC}(A,B)=\frac{|\mathrm{LL}|}{|\mathrm{LL}|+|\mathrm{LO}|+|\mathrm{OL}|} \tag{11}$$

$$\mathrm{LL}=\mathrm{leader}(A)\cap\mathrm{leader}(B) \tag{12}$$

$$\mathrm{LO}=\mathrm{leader}(A)\cap\mathrm{ordinary}(B) \tag{13}$$

$$\mathrm{OL}=\mathrm{ordinary}(A)\cap\mathrm{leader}(B) \tag{14}$$

式中,$\mathrm{leader}(X)$表示 X 话题中的意见领袖集合,$\mathrm{ordinary}(X)$表示 X 话题中的普通用户集合。

3.4 实验结果

3.4.1 不同种子集的扩展核心率对比

为了验证 EIC 算法的有效性,将 EIC 算法与 OLR、MR 以及 PR 算法挖掘结果的扩展核心率进行对比。经实验分析,当 ECR 中的参数 ε 取值在 0.4 以上,φ 取值在 0.7 以上时,扩展核心率较为稳定,因此实验中 ε、φ 分别设置为 0.4、0.7。将种子节点数量依次从 5 递增到 30,对比不同算法挖掘出的种子集的平均扩展核心率。实验结果如图 3~6 所示。

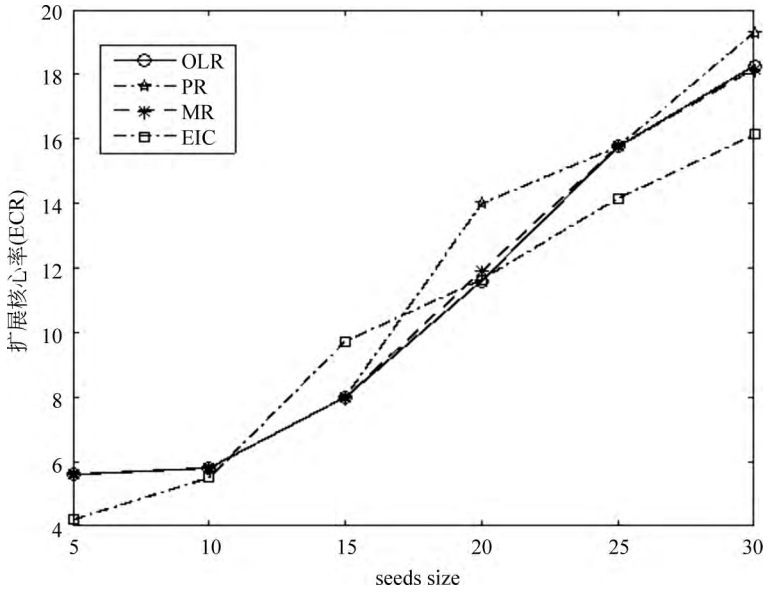

图 3 "两会房价"种子集扩展核心率曲线

从实验结果来看,EIC 算法的扩展核心率相对较小,尤其是种子节点数大于 20 时,远小于其他算法的扩展核心率。文中为方便计算,将所有节点的扩展核心率由高到低排序,重新赋值为 $1\sim N$(N 表示节点数),进而扩展核心率越低,代表节点属性权重比率以及网络核心率越高,意见领袖能力越强,因而该算法在扩展核心率指标上优于其他算法。此外,EIC 算法在选取意见领袖时,不仅考虑了节点的个体属性、网络结构特征,而且考虑了节点的信息传播能力;而 PR 算法仅考虑了网络链接关系,OLR 算法和 MR 算法仅考虑了交互特征以及部分节点属性特征。

EIC 算法不仅扩展核心率较低,其稳定性也优于其他三种算法。图 3~6 中,EIC 算法的扩展核心率变化范围始终最小,曲线过渡平缓,稳定性相对较好。而图 4 中的 PR 算法在种子节点数为 25~30 时,其扩展核心率急剧增大,数据质量不高,且曲线稳定性较差;从图 5 中可以看出,虽然在种子节点数为 15 附近时,其他三种算法的扩展核心率小于 EIC 算法,但曲线整体波动较大,稳定性相比 EIC 算法较差;在图 6 中,其他三种算法的扩展核心率不仅数值较大,而且曲线的波动幅度也很大,稳定性相对

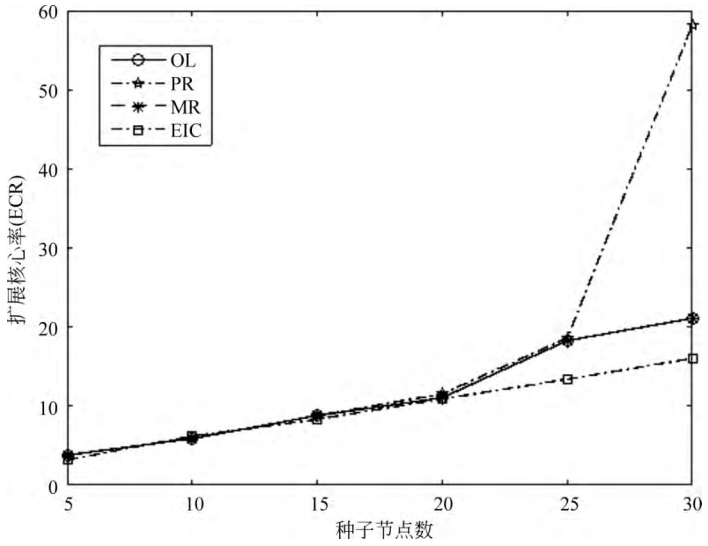

图 4　“两会雾霾”种子集扩展核心率曲线

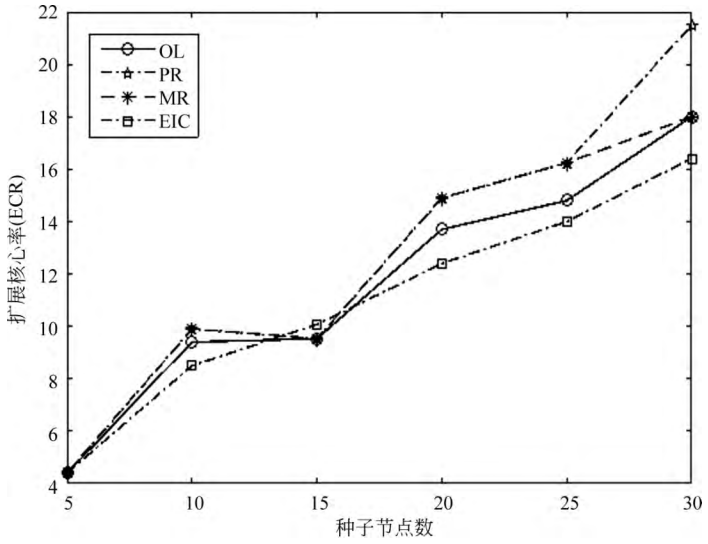

图 5　“嫦娥三号”种子集扩展核心率曲线

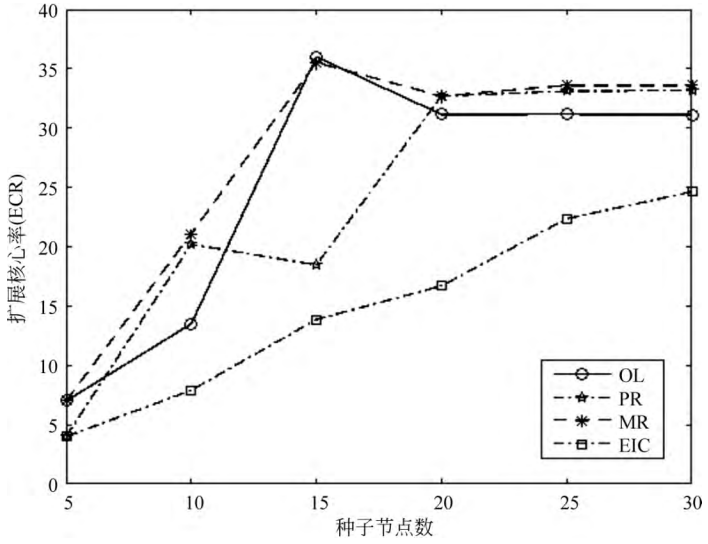

图 6　“单独二胎”种子集扩展核心率曲线

较差，而EIC算法的扩展核心率曲线始终在较低位置平缓变化，稳定性较好。

因此，在统计分布特征不同的多个话题集下，EIC算法挖掘出的意见领袖在扩展核心率上优于其他拓扑结构类算法，且具有较好的稳定性。

3.4.2 不同种子集的影响传播范围对比

为进一步研究分析EIC算法，将种子节点数量依次从5递增到30，对比不同算法产生的种子集的影响传播范围。实验结果如图7～10所示。

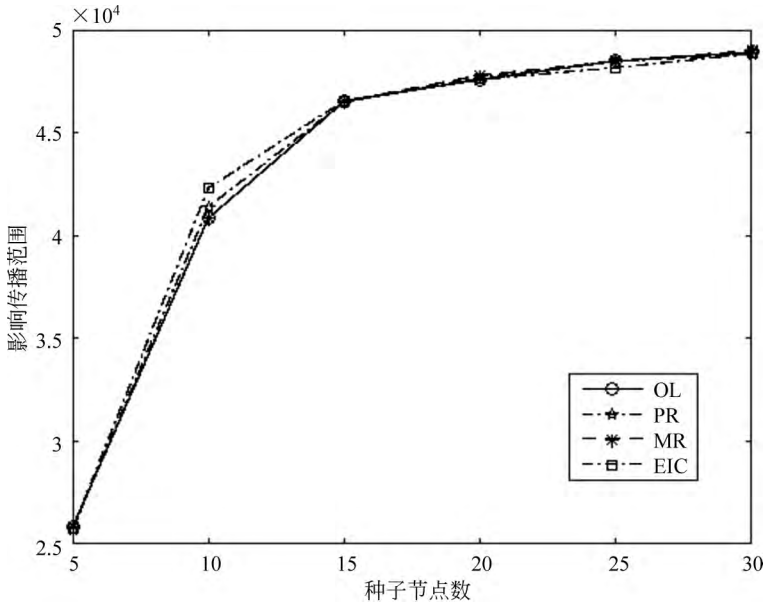

图7 “两会房价”种子集影响传播范围曲线

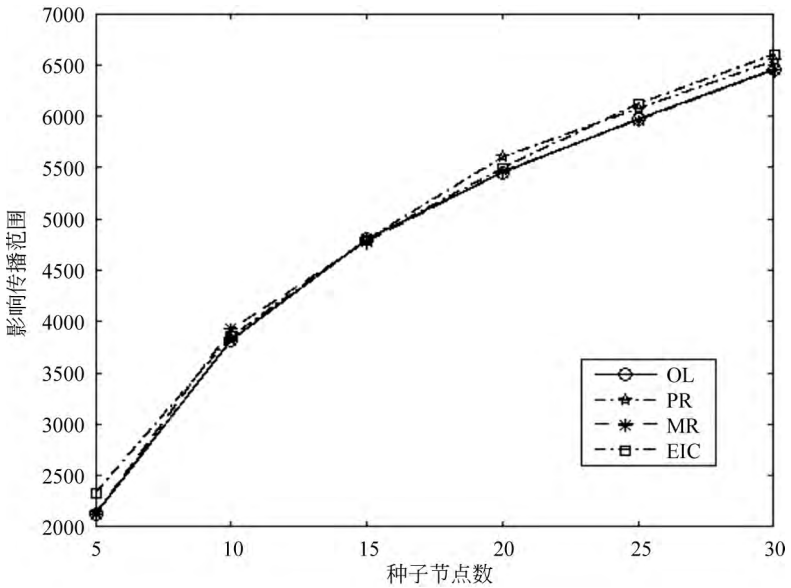

图8 “两会雾霾”种子集影响传播范围曲线

从图7～10中可以看出，EIC算法和其他三种算法的影响传播范围大致相同，但在某些特定的种子节点数下（如图7、图10中种子节点数10附近；图8、图9中种子节点数30附近）略高于其他拓扑结构类算法，这是因为EIC算法在选取意见领袖时，不仅考虑了个体属性、网络结构特征，还考虑了用户节点的信息传播能力。

从图8可以看出，EIC算法的影响传播范围相对较大，信息扩散能力较强，并且始终超出OLR算法的影响传播范围；在图9中，EIC算法的传播范围虽然略低于PR算法，但这是因为“嫦娥三号”数据集的平均路径长度相对较小，导致节点间的信息传

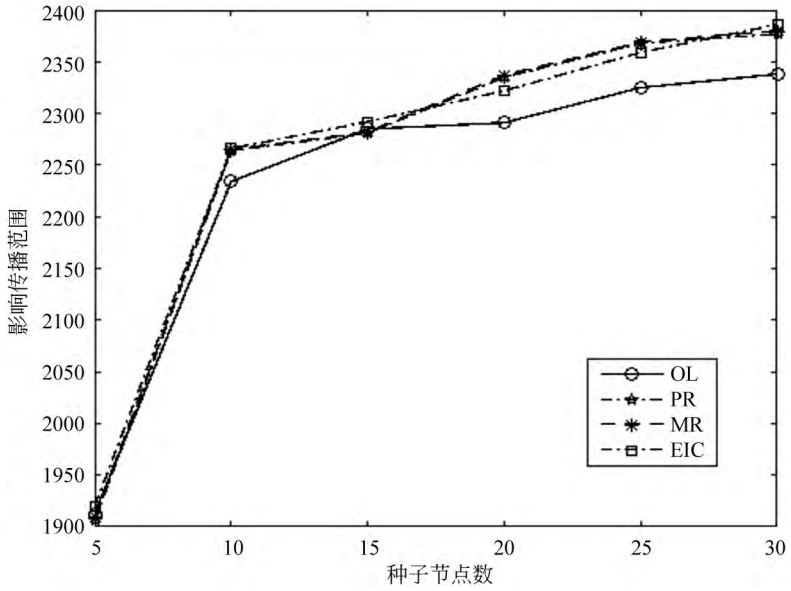

图 9　“嫦娥三号”种子集影响传播范围曲线

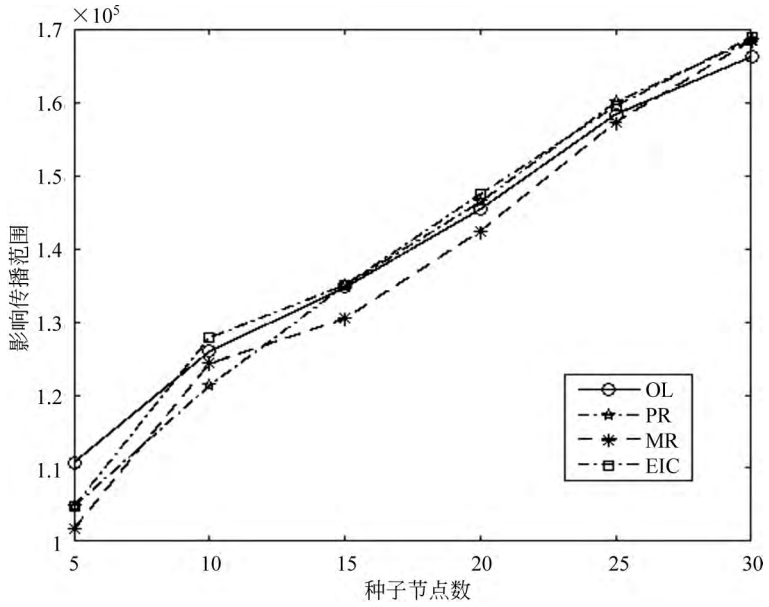

图 10　“单独二胎”种子集影响传播范围曲线

递速度相对较大，因而越有利于拓扑结构类算法，不具有普遍性；在图 10 中，当种子节点数大于 10 时，EIC 算法的影响传播范围基本处于最大，优于其他三种算法。

因此，在统计分布特征不同的多个话题集下，EIC 挖掘出的意见领袖在传播范围上整体优于其他拓扑结构类算法，尤其话题网络的平均路径长度越大，这种优势越明显。

3.4.3　意见领袖的话题特征分析

表 6 展示了 EIC 算法在四个不同的话题下挖掘出的排名前十的意见领袖。从表中可以看出，“人民日报”“央视新闻”“南方都市报”等用户在多个话题中都为意见领袖，但多数用户仅为某个特定话题中的意见领袖。

表 6　EIC 算法在不同话题下的前十意见领袖

排名	两会房价	两会雾霾	嫦娥三号	单独二胎
1	人民日报	国家电网	央视新闻	央视新闻
2	央视新闻	新浪财经	搜狐视频	人民日报
3	十三点半的 kings	杨澜	新京报	南方都市报

续表

排名	两会房价	两会雾霾	嫦娥三号	单独二胎
4	洗屌少女	中国之声	嫦娥三号	协和章蓉娅
5	唐僧僧僧僧僧	南方都市报	长沙吃喝玩乐情报	头条新闻
6	摆古论今	最高人民法院	都市快报	人民网
7	郭敬明	挺么么	中国之声	环球时报
8	烧伤超人阿宝	北京环境监测	安徽卫视	新浪财经
9	崔永元	华尔街日报中文网	国家海洋预报台	21世纪经济报道
10	新浪江苏	知乎	BigBang情	广州日报

对意见领袖一致性检测结果如表7所示，意见领袖一致性相对较小，只有少量用户可以成为多个话题中的意见领袖，大部分意见领袖只出现在特定话题中，各个话题间的意见领袖相对独立。进而说明意见领袖一致性越小，话题依赖程度越大，意见领袖的话题特征越明显。因此，在挖掘微博平台的意见领袖时，话题特征也是不可忽略的因素。

表7 意见领袖一致性检测

话题	两会房价	两会雾霾	嫦娥三号	单独二胎
两会房价	—	0	0.17	0.18
两会雾霾	0	—	0.20	0.22
嫦娥三号	0.17	0.20	—	0.33
单独二胎	0.18	0.22	0.33	—

4 总结

社交网络中意见领袖的挖掘对于舆情监控、信息扩散等方面具有重要的应用价值。用户在社交网络中以话题、兴趣爱好等形式迅速聚合为群体，实现群体交互关系。本文以用户交互关系为基础，将分散的用户个体聚合到大型、复杂的加权传播网络中，提出了一个基于扩展独立级联模型，并融入网络结构特征、个体属性和行为特征的意见领袖挖掘模型。本文在真实的微博数据集上对模型进行验证，实验结果表明基于该模型挖掘出的意见领袖质量更高。

后续工作可以考虑以下两点：第一，在本文融合多种属性构建的模型中未考虑微博文本内容以及文本情感倾向，下一步可以增加属性维度，完善传播模型；第二，影响传播概率采用的是最简单的线性加权，未来可以通过机器学习相关算法优化参数，采用逻辑回归等方法进行信息融合，从而提高微博平台下挖掘意见领袖的精确度。

参考文献

[1] Vespignani A. Modeling Dynamical Processes in Complex Socio-technical Systems[J]. Nature Physics, 2011, 8(1): 32-39.

[2] Bond R M, Fariss C J, Jones J J, et al. A 61-million-person experiment in social influence and political mobilization. Nature, 2012, 489(7415): 295-298.

[3] Asur S, Huberman BA, Szabo G, et al. Trends in social media: Persistence and decay[J]. Social Science Electronic Publishing, 2011.

[4] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]// International Conference on Weblogs and Social Media, Icwsm 2010, Washington, Dc, Usa, May. DBLP, 2010.

[5] 许丹青，刘奕群，张敏，等. 基于在线社会网络的用户影响力研究[J]. 中文信息学报, 2016, 30(2): 83-89.

[6] Lin K C, Wu S H, Chen L P, et al. Finding the Key Users in Facebook Fan Pages via a Clustering Approach[C]//Proceedings of IEEE International Conference on Information Reuse and Integration. IEEE, 2015: 556-561.

[7] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[J]. Wsdm, 2010: 261-270.

[8] 吴岘辉，张晖，杨春明，等. 一种话题相关的微博意见领袖挖掘算法[J]. 小型微型计算机系统, 2014, 35(10): 2296-2301.

[9] Yang L, Qiao Y, Liu Z, et al. Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm[J]. Soft Computing, 2016: 1-12.

[10] 朱玉婷，李雷，施化吉，等. 社会网络中基于主题的影响力最大化算法[J]. 计算机应用研究, 2016, 33(12): 3611-3614.

[11] Jendoubi S, Martin A, Liétard L, et al. Two Evidential Data Based Models for Influence Maximization in Twitter[J]. Knowledge-Based Systems, 2017, 121: 58-70.

[12] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003: 137-146.

（下转第146页）

1405.4053,2014.
[18] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,6(40):1-23.
[19] 李彦冬,郝宗波,雷航.卷积神经网络研究综述[J].计算机应用,2016,39(09):2508-2515,2565.
[20] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
[21] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.
[22] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences[C]// Proceedings of Advances in Neural Information Processing Systems. USA: MIT Press, 2014: 2042-2050.
[23] Bengio Y, Schwenk H, Senécal J, et al. Neural probabilistic language models[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
[24] Collobert R, Weston J, Bottou L, et al. Natural language processing(almost)from scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.
[25] Kiros R, Zhu Y, Salakhutdinov R, et al. Skip-thought vectors[C]// Proceedings of International Conference on Neural Information Processing Systems. USA: MIT Press, 2015: 3294-3302.
[26] Lin C Y, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, USA: Stroudsburg, 2004.

穆婉青(1993—),硕士研究生,主要研究领域为自然语言处理。
E-mail: 948611255@qq.com

王素格(1964—),通信作者,博士,教授,主要研究领域为自然语言处理,文本情感分析。
E-mail: wsg@sxu.edu.cn

廖健(1990—),博士研究生,主要研究领域为文本情感分析。
E-mail: liaojian_iter@163.com

(上接第138页)

[13] Saaty T L. Basic theory of the analytic hierarchy process: how to make a decision[J]. Revista De La Real Academia De Ciencias Exactas Físicas Y Naturales, 2007, 93(4): 395-423.
[14] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 420-429
[15] LinY, LiH, Liu X, et al. Hot topic propagation model and opinion leader identifying model in microblog network[J]. Abstract andApplied Analysis, 2013, 36(2): 360-367.
[16] 刘志明,刘鲁.微博网络舆情中的意见领袖识别及分析[J].系统工程,2011,29(6):8-16.

张米(1992—),硕士研究生,主要研究领域为数据挖掘、社交网络。
E-mail: zhang_mi66@163.com

杨春明(1980—),硕士,副教授,主要研究领域为文本挖掘、知识工程。
E-mail: yangchunming@swust.edu.cn

张晖(1972—),通信作者,博士,教授,主要研究领域为文本挖掘、知识工程。
E-mail: zhanghui@swust.edu.cn