基于用户行为网络的微博意见领袖挖掘算法*

吴岘辉[1a], 张 晖[1b†], 赵旭剑[1a], 李 波[1a,2], 杨春明[1a]

(1. 西南科技大学 a. 计算机科学与技术学院;b. 教育信息化推进办公室, 四川 绵阳 621010; 2. 中国科学技术大学 计算机科学与技术学院, 合肥 230027)

摘 要: 微博意见领袖挖掘中通常单独考虑用户属性、网络结构或交互信息等特征,对这些特征之间的关系及微博信息的话题特征考虑较少。针对上述问题,提出了一种基于用户行为网络的微博意见领袖挖掘算法 TopicLeaderRank。该算法利用微博用户的内容属性和社交属性,并结合用户在特定话题中的交互信息构建用户行为网络,然后利用 PageRank 算法的投票思想,同时考虑网络中节点权重和边权重对投票的影响来挖掘意见领袖。在新浪微博三个话题数据集上的实验结果表明,该算法是有效的,在覆盖度和核心率指标上的值高于用户权重排序和 Microblog-Rank 算法,在人工评价上的表现也优于这两种算法。

关键词: 意见领袖; 微博; 话题; PageRank

中图分类号: TP391.1; TP301.6 **文献标志码:** A **文章编号:** 1001-3695(2015)09-2678-06

doi:10.3969/j.issn.1001-3695.2015.09.028

Mining algorithm of microblogging opinion leaders based on user-behavior network

Wu Xianhui[1a], Zhang Hui[1b†], Zhao Xujian[1a], Li Bo[1a,2], Yang Chunming[1a]

(1. *a. School of Computer Science & Technology, b. Educational Informationization Office, Southwest University of Science & Technology, Mianyang Sichuan* 621010, *China*; 2. *School of Computer Science & Technology, University of Science & Technology of China, Hefei* 230027, *China*)

Abstract: In the mining of microblogging opinion leaders, features were always separately considered, such as users' attributes, network structures and mutual information. Researchers rarely considered the relationship among these features, as well as the topic features of information in microbloggings. Aimed at above problems, this paper proposed an algorithm, named TopicLeaderRank, for opinion leaders mining in microbloggings based on the user-behavior network. First, the algorithm built a user-behavior network by using users' content and social attributes and mutual information about a specific topic. Then it adopted the thought of voting in the PageRank algorithm to find opinion leaders. It also considered the influence from the weights of nodes and edges in the use-behavior network on the voting. Experiments based on three data sets of Sina MicroBlog show that the algorithm is efficient. It gets higher scores on indicators of coverage and coreratio than the sorting algorithm by user's weights and Microblog-Rank algorithm. It also performs better than the other two algorithms in the artificial evaluation.

Key words: opinion leader; microblogging; topic; PageRank

0 引言

意见领袖通常指与周围的人存在较多联系,同时拥有突出教育水平和身份地位从而能影响其追随者的人[1]。由于微博具有信息传播快、时效性强、应用广泛等特点,微博中的意见领袖能对整个网络的信息传播和观点形成产生重要影响,他们的意见在很大程度上反映了整个网络的意见。因此,微博中的意见领袖挖掘研究,有助于更好地理解公众意见的形成,对于网络舆情监控有重要意义,同时有利于微博中网络营销和高质量好友推荐等应用的推广。

意见领袖往往是拥有较大的社会影响力的用户[2]。网络用户的社会影响力主要体现在用户属性、网络结构和信息交互三个方面,因此多数研究者分别从这三个方面来研究网络意见领袖的挖掘。在微博中,用户属性和信息交互体现了用户的活跃性,网络结构能反映用户的网络中心性。活跃性和网络中心性是意见领袖的必要特征,活跃性保证了意见领袖对其他用户影响力的来源,网络中心性则提供了意见领袖对他人施加影响的途径。因此,综合考虑用户属性、网络结构和信息交互三方面的信息,应该能够更有效地挖掘微博中的意见领袖。

此外,微博中的讨论具有显著的话题特征,而用户在不同的话题中往往具备不同的影响力[3]。用户在自己擅长或感兴趣的话题中参与度更高,表现更活跃,因此往往具有更强的影响力。然而,已有研究在挖掘意见领袖时,大多从全局或者知识领域的角度去衡量用户的影响力[4~7],忽视了意见领袖的话题领域性。

针对以上问题,本文从微博用户的话题特征入手,全面分

收稿日期: 2014-06-30; **修回日期:** 2014-08-06 **基金项目:** 四川省教育厅基金资助项目(14ZB0113);西南科技大学博士基金资助项目(12zx7116)

作者简介: 吴岘辉(1989-),男,四川大竹人,硕士,主要研究方向为数据挖掘、社交网络;张晖(1972-),男(通信作者),教授,博士,主要研究方向为文本挖掘、知识工程(zhanghui@swust.edu.cn);赵旭剑(1984-),男,讲师,博士,主要研究方向为中文信息处理、Web 信息检索;李波(1977-),男,讲师,博士,主要研究方向为信息过滤、信息安全;杨春明(1980-),男,讲师,硕士,主要研究方向为文本挖掘、知识工程.

析用户属性以及用户关系,构建特定话题下的用户行为网络,通过网络分析方法挖掘话题意见领袖。与已有工作相比,本文的主要贡献在于:

a)建立了一个面向微博话题的用户行为网络,从用户属性(静态特征)和用户关系(动态特征)两方面挖掘用户的影响力,综合考虑了用户在特定话题领域的领袖作用。

b)提出了一种基于改进的 PageRank 算法的话题意见领袖挖掘算法,针对特定话题的用户行为网络,在传统 PageRank 算法的基础上扩展网络属性特征,从基础影响力以及话题领域影响力两方面挖掘意见领袖,提高了话题意见领袖挖掘结果的准确性。

1　相关工作

目前关于网络意见领袖挖掘的研究,主要分为三类方法:a)用户属性分析法;b)信息交互分析法;c)网络结构分析法。

基于用户属性的相关工作中,丁雪峰等人[8]采用构造用户的属性矩阵并进行综合权重排序的方法,祝帅等人[9]设计了基于贝叶斯信息增益最大化准则的 X-means 迭代聚类筛选模型,Zhang 等人[10]通过马尔可夫网络对用户的内在属性、社交属性和内容属性的相关性进行分析来挖掘意见领袖。

信息交互分析法主要通过分析用户所发信息的影响力及其传播特性来反映用户的影响力,从而挖掘意见领袖。Agarwal 等人[11]综合考虑了博文的引用数量、评论数量、新颖程度和内容长度评价帖子的影响力;Li 等人[1]综合考虑了博文的数量和质量来挖掘商品评价中的意见领袖;樊兴华等人[12]通过改进影响力扩散模型提出了衡量论坛帖子影响力的影响力概率扩散模型。

相比以上两类方法,也有不少学者提出利用网络结构挖掘意见领袖的策略:根据用户间的关系建立图模型,然后以网络结构分析为主,通过衡量用户在网络中的重要性(中心性)来挖掘意见领袖。根据分析方法的不同,这类研究可以分成两个方向:a)利用复杂网络理论来分析用户的网络中心性[13];b)通过 PageRank[14]等网络拓扑分析法及它们的扩展算法来分析用户的网络重要性,如考虑博文本身的新颖性和博文之间的链接关系的 InfluenceRank 算法[6]、考虑论坛回复的情感极性的 LeaderRank 算法[15]、考虑用户关注关系和话题相似性的 TwitterRank 算法[16]及考虑用户评论关系的 Microblog-Rank 算法[17]。肖宇等人[18]的研究则结合了这两个方向来挖掘意见领袖。基于网络结构分析的方法能在很大程度上反映用户的网络影响力,该类中的部分算法在构建图模型也考虑了一定的交互信息,因此在挖掘意见领袖时能得到较好的效果,也是目前的主流研究方法。

以上三类方法是从三个互补的方面去挖掘意见领袖,因此单独考虑任意一个方面都不完整。在微博中,目前还没有从话题的角度出发并综合考虑用户属性、网络结构和信息交互的意见领袖挖掘算法。在上面提到的研究中,只有 TwitterRank 和 Microblog-Rank 算法考虑了话题特征。尹衍腾等人[19]提出了一种基于用户关系与属性的微博意见领袖挖掘方法,但该方法仅仅是把基于用户属性的方法和基于网络结构的方法简单相加,并没有把两种方法真正融合起来,同时未考虑用户间的信息交互情况。冯时等人[20]提出了一种基于 LDA 模型的意见领袖挖掘方法,该方法先通过 LDA 模型和 HowNet 知识库来获取话题,然后通过层次分析法综合考虑用户的属性和用户发表信息的影响力来获取意见领袖,但未考虑用户间的网络结构。

与本文工作最为相似的研究是文献[17]提出的意见领袖挖掘算法 Microblog-Rank,但它与本文提出的意见领袖挖掘算法存在明显的区别。首先 Microblog-Rank 算法在构建交互图模型时未考虑边的方向(影响力的方向),仅考虑了用户间的微博评论数,同时通过下面的分析可以看出该算法实际上未用到用户的属性特征。Microblog-Rank 算法设置图模型中的边权重 w_{ij} 如式(1)所示。其中 $p(i)$ 表示用户属性权重,N_{ij} 表示用户之间的评论数,其迭代式如式(2)所示。将式(1)代入式(2)可得该算法的最终迭代公式如式(3)所示,从式(3)可以看出 Microblog-Rank 算法实际上并未考虑用户的属性,而只考虑了用户间的微博评论数。本文提出的算法不仅考虑了用户间影响力的方向,同时对用户属性以及用户之间的微博评论和转发关系进行了综合考虑,比 Microblog-Rank 算法考虑得更加细致全面。

$$w_{ij} = p(i) \times N_{ij} \tag{1}$$

$$MR(i) = (1-a) + a \sum_{\text{edge}\langle v_j, v_i\rangle} \frac{MR(j) w_{ji}}{\sum_{\text{edge}\langle v_j, v_k\rangle} w_{jk}} \tag{2}$$

$$MR(i) = (1-a) + a \sum_{\text{edge}\langle v_j, v_i\rangle} \frac{MR(j) N_{ji}}{\sum_{\text{edge}\langle v_j, v_k\rangle} N_{jk}} \tag{3}$$

2　意见领袖挖掘算法

2.1　用户行为网络

将微博中的用户看做节点,用户之间的联系看成边,整个微博网络则构成一个大的图,而话题相关的用户及用户间的交互关系就是其中的一个子图。该子图可被看做特定话题下的用户行为网络,采用有向加权图 $G=(V,E,P,W)$ 表示,其中 V 表示话题相关的用户节点集合,E 表示用户间的有向边集合,P 表示节点权重集合,W 表示边权重集合。下面分步构建用户行为网络。

首先构建用户行为网络的节点和边。当微博用户 v_i 参与到了特定话题的讨论中,具体表现为发表、转发或者评论了话题相关的微博,则将该用户加入到特定话题图模型的节点集合中,即 $v_i \in V$。在特定话题下,用户之间的交互关系比用户间的关注关系更能体现用户之间的影响情况,因此采用用户之间的交互关系在图模型中建立边,主要考虑用户间的转发和评论情况。考虑到用户之间的影响是有方向性的,即当用户 a 在转发或者评论他人的微博时,通常表示用户 a 受到他人的影响。因此,当用户 v_i 转发或者评论了用户 v_j 话题相关的微博时,则创建一条从 v_i 指向 v_j 的有向边 $e_{ij} \in E$。

在构建节点权重集合和边权重集合时,对用户属性和用户间交互属性进行了综合考虑,具体考虑的权重影响因素如表 1 所示。从表中可以看出,用户属性由内容属性(信息量)和社交属性(与周围用户的联系紧密度)组成,内容属性又分为话题相关和话题无关两类,而交互属性仅包括话题相关的内容属性。

表 1　节点权重和边权重的影响因素

属性	详细指标
用户属性	总的微博数
	话题相关的微博数、评论数和转发数、关注数、粉丝数和相互关注数
交互属性	话题相关的评论数和转发数

用户 v_i 的权重 $p_i \in P$,与用户属性有关,具体计算公式可

表示为式(4)。其中 N_1 表示用户的总微博数;$N_2 \sim N_4$ 分别表示话题相关的微博数、评论数和转发数;$N_5 \sim N_7$ 分别表示用户的关注数、粉丝数以及相互关注数;$\beta_1 \sim \beta_7$ 分别表示相应子属性占的权重比例,它们的和为 1。边 e_{ij} 的权重 $w_{ij} \in W$,与用户 v_i 对用户 v_j 话题相关的微博转发数和评论数有关,具体按式(5)计算,其中 N_8 和 N_9 表示相应的转发数和评论数,λ 表示转发数占的权重比例。

$$p_i = \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \beta_4 N_4 + \beta_5 N_5 + \beta_6 N_6 + \beta_7 N_7 \quad (4)$$

$$w_{ij} = \lambda N_8 + (1-\lambda) N_9 \quad (5)$$

2.2 TopicLeaderRank 算法

在构建好特定话题下的用户行为网络之后,挖掘意见领袖就转换为寻找图中心节点的问题。传统的图中心节点挖掘算法,例如复杂网络中的中心性分析法和 PageRank 算法,均未考虑到图中的节点权重和边权重,因此并不适用于本文构建的用户行为网络。本文通过对 PageRank 算法进行扩展,提出了基于用户行为网络的话题意见领袖挖掘算法 TopicLeaderRank,该算法除了考虑网络拓扑结构,还融入了对节点权重和边权重的考虑。

TopicLeaderRank 算法采用了 PageRank 算法的迭代投票思想,并扩展了 PageRank 算法的投票策略。PageRank 算法采用的是链接投票的策略,即边 e_{ij} 表示节点 v_i 对节点 v_j 的一次投票,节点在投票时会将自身的 PageRank 权重按出度平均分配到指向的节点,因此该算法得出的节点重要性与节点的度密切相关,特别是节点的入度。TopicLeaderRank 在投票时采用了网络拓扑与权重相结合的策略,这里的权重是指节点权重和边权重。

TopicLeaderRank 算法还是按照网络拓扑来投票,一条有向边表示一次投票。因为投票代表了被影响关系,所以在本文构建的用户行为网络中,边是从主动评论(或者转发)的用户指向被评论(或者被转发)的用户。但是,节点在投票时不再对所有链接到的节点一视同仁,而要考虑边权重。边权重代表了特定话题下用户之间的交互程度,权重越大,用户之间的交互越密切,用户间的影响也越大。因此,节点在投票时,根据边的权重来分配自身的重要性权重(用 TopicLeaderRank 值表示,简称 TLR 值)更合理一些,即投票本身有权重,用户向对其影响更大的用户投票权重也更大。另外,投票权重除了与边权重有关,还与投票用户自身的权重有关。用户节点的权重综合考虑了用户的活跃性、社交关系紧密性和话题相关性,在一定程度上代表了用户的重要性,而重要性越大的用户往往话语权也越大,投票也越有分量。因此,投票权重与节点权重和边权重都成正比。

按以上分析,TopicLeaderRank 算法的迭代公式可表示为式(6),其中 $\pi^{(k)\mathrm{T}}$ 和 $\pi^{(k+1)\mathrm{T}}$ 分别表示更新前后的所有节点的 TLR 值向量,α 为阻尼系数,e^{T} 表示单位行向量,V 为投票矩阵,其中的元素 v_{ij} 表示投票权重,按式(7)计算。阻尼系数 α 用来解决真实图的稀疏问题,保证算法的正确收敛。另外,在图模型中往往存在一些出度为 0 的节点,这类节点被称为死节点。死节点的存在会影响算法效果,因此算法在使用迭代公式前需要先反复去掉图中的死节点,然后按照迭代式计算剩余节点的 TLR 值,收敛后再根据网络拓扑反推死节点的 TLR 值。最终,TLR 值排名靠前的用户被当做意见领袖。

$$\pi^{(k+1)\mathrm{T}} = (1-\alpha) e^{\mathrm{T}} + \alpha \pi^{(k)\mathrm{T}} V \quad (6)$$

$$v_{ij} = \begin{cases} 0, e_{ij} \notin E \\ \dfrac{p_i w_{ij}}{\sum_{e_{ik} \in E} w_{ik}}, e_{ij} \in E \end{cases} \quad (7)$$

TopicLeaderRank 算法是一个迭代的算法,令算法收敛后的解为 π^*,则 π^* 满足式(8)。将式(8)看做一个向量矩阵方程,并令矩阵 A 如式(9)所示,其中 E 为与 V 同阶的单位矩阵。因此当 A 可逆时,算法有唯一解,如式(10)所示。矩阵可逆的条件并不苛刻,当矩阵对应的所有行向量是线性无关组时,矩阵可逆。值得注意的是,矩阵 V 的每一行代表图模型中某一个节点对其他节点的投票权重向量,这与节点本身的权重以及该节点与其他节点的交互情况密切相关,所以矩阵 V 的每一行都是不相关的,按照式(9)的计算,矩阵 A 的每一行也没有相关性,所以矩阵 A 存在逆矩阵。因此,TopicLeaderRank 算法在一般情况下都存在唯一解。仅当图中存在两个节点对其他所有节点的投票权重对应成比例的情况下,矩阵 V 不存在逆矩阵,这时算法没有唯一解。由于每个节点的属性存在差异,同时每个节点与邻居节点的交互情况也各异,所以这种特殊情况极少出现。

$$\pi^* = (1-\alpha) e^{\mathrm{T}} + \alpha \pi^* V \quad (8)$$

$$A = E - \alpha V \quad (9)$$

$$\pi^* = (1-\alpha) e^{\mathrm{T}} A^{-1} \quad (10)$$

3 实验

3.1 实验数据与设置

本文采用新浪微博中近期三个热门话题的数据集作为实验数据,包括最近关于“单独二胎”政策的讨论以及 2014 年两会期间关于房价和雾霾的讨论,详细信息如表 2 所示。为了保证微博用户和微博信息的话题相关性,数据的获取包含两个阶段:a)从科研数据共享平台数据堂(http://www.datatang.com/)下载其他用户整理后上传的相关话题的原始微博数据;b)通过新浪微博开放平台的官方 API 获取这些微博的转发、评论及涉及的用户信息。

表 2 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
单独二胎	40819	274096	254850	354193	2013.11.1 ~ 2014.5.17
两会房价	13111	35880	35105	67646	2014.3.3 ~ 2014.4.19
两会雾霾	6997	7731	7527	17151	2014.3.3 ~ 2014.4.22

在本文的实验中,将 TopicLeaderRank 算法与仅考虑用户属性的节点权重排序算法和仅考虑用户间交互关系的 Microblog-Rank[17] 算法进行了对比实验。其中节点权重排序算法是本文设置的一个基准算法,它仅仅对于所有参与话题的用户按照节点权重从大到小排序,然后将排名靠前的用户当做意见领袖。由于原始 Microblog-Rank 算法仅考虑用户间的评论关系,而本文构建的用户行为网络同时考虑了用户间的评论和转发关系,所以在实验时对 Microblog-Rank 算法进行了细微的扩展,即同时考虑用户间的评论和转发关系建立边,另外式(1)中 N_{ij} 取本文图模型中的 w_{ij} 的值。

在上述三个算法中涉及的参数取值如表 3 所示。在这些参数中,对算法有较大影响的是阻尼系数 α,它是算法有效性与收敛速率之间的平衡参数。由于 TopicLeaderRank 和 Microblog-Rank 算法都是 PageRank 算法的扩展算法,所以它们的 α 参数均取 0.85,与 PageRank 算法常用的阻尼系数一致。表 3

中的其他参数对用户行为网络中的节点权重和边权重有直接影响,因此对节点权重排序算法的影响较大,但对其他两种算法影响较小,因为其他两种算法主要依赖于用户行为网络的拓扑结构。同时,这些参数均是用户行为网络中相关子属性的比例参数。因此,这些参数相对次要,在实验中仅利用主观经验对它们进行人为设置:在计算用户节点权重时,用户的微博数和粉丝数更能体现用户的活跃性以及对他人的影响力,因此相应的比例参数取 0.2,其他的子属性的比例参数则取 0.1;在计算边权重时,用户的评论数和转发数重要性相差不大,因此比例参数 λ 取 0.5。

表 3　参数取值情况

公式	参数取值
式(2)	$\alpha = 0.85$
式(4)	$\beta_1 = \beta_2 = \beta_6 = 0.2$, $\beta_3 = \beta_4 = \beta_5 = \beta_7 = 0.1$
式(5)	$\lambda = 0.5$
式(6)	$\alpha = 0.85$

3.2　评价方法

目前还没有一个公认的网络意见领袖评价标准。在本文的实验中,采用了覆盖度(coverage)[6]、核心率(coreratio)[15]以及人工评价三种方法来评价意见领袖。

覆盖度是从用户交互形成的网络拓扑的角度出发,通过计算影响的用户数来衡量意见领袖能力。在本文构建的网络图模型中,一个用户影响的用户集合为通过图中的有向边可以到达该用户的用户群。根据计算方式的不同,覆盖度分为单步覆盖度(one-step coverage)和全路径覆盖度(all-path coverage)。单步覆盖度仅考虑用户直接影响的用户数,即不考虑影响力的传播;全路径覆盖度则考虑用户直接或者间接影响的全部用户数,即认为影响力会通过边一直传播下去,而且不会减弱和消失。但在真实的社交网络中,影响力不会局限于和意见领袖直接交互的用户,也无法沿着边一直传递下去而不减弱和消失。因此,本文实验中覆盖度的计算是取单步覆盖度和全路径覆盖度的平均值。

核心率也是从用户交互的角度来衡量意见领袖的能力。在本文构建的网络图模型中,用户的核心率是指其他用户对该用户的交互信息量(边权重)占总的交互信息量的比率。用户的核心率由其他用户对该用户话题相关的微博转发数和评论数决定的。

人工评价是指通过人为分析意见领袖的相关特征来判断其意见领袖能力的评价方法。本文在对意见领袖进行人工评价时,综合考虑了意见领袖的本身属性和网络特性,本身属性包括信息量、话题参与度、与其他用户的亲密度(关注数、粉丝数和相互关注数)以及用户在现实社会的权威度,网络特性主要指意见领袖在网络图模型中的入度和出度。

3.3　实验结果

本文比较了三种算法在三个话题中找出的意见领袖的覆盖度和核心率。图 1 ~ 3 展示了意见领袖值排名前 20 的用户的覆盖度情况,图 4 ~ 6 展示了意见领袖值排名前 1% 的用户核心率情况。从这些图中可以看出,在覆盖率和核心率这两个评价指标上,节点权重排序算法的效果是最差的,且与其他两种算法的效果有很大的差距。这是因为覆盖率与核心率指标都是从用户交互的角度来评价意见领袖,而节点权重排序算法仅仅考虑了用户的属性权重,而忽略了用户之间的交互情况。TopicLeaderRank 算法在三个话题下的覆盖率和核心率都要高于其他两种算法,特别是在“单独二胎”和“两会雾霾”两个话题下。同时 TopicLeaderRank 算法得到的前 20 个用户的覆盖率平均在 40% 以上,前 1% 的用户核心率平均在 80% 以上,对应结果曲线的增长率存在明显的下载趋势,即排名靠前的用户拥有更大的覆盖率和核心率。这说明 TopicLeaderRank 算法的思路是正确的,同时要优于其他两种算法,虽然覆盖度与核心率都未涉及用户的属性,但是同时考虑用户属性和用户交互的 TopicLeaderRank 算法要比仅考虑用户交互的 Microblog-Rank 算法更有效。

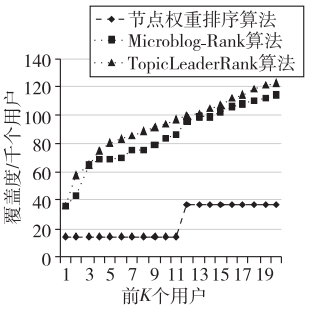

图 1　三种算法在“单独二胎”话题中的覆盖度

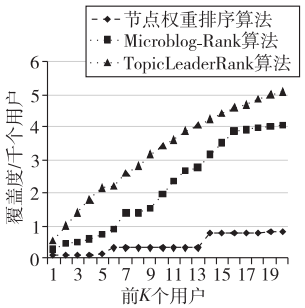

图 2　三种算法在“两会雾霾”话题中的覆盖度

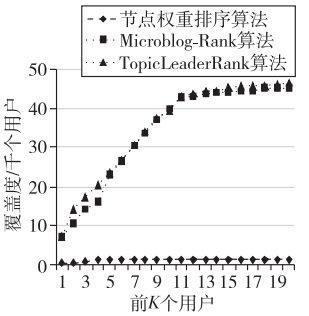

图 3　三种算法在“两会房价”话题中的覆盖度

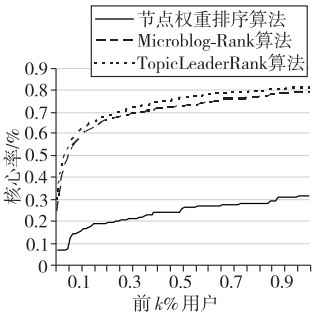

图 4　三种算法在“单独二胎”话题下的核心率

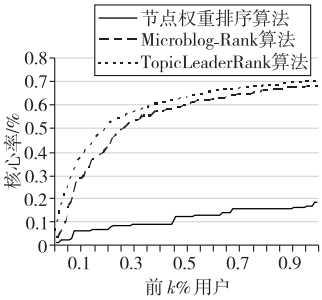

图 5　三种算法在“两会雾霾”话题下的核心率

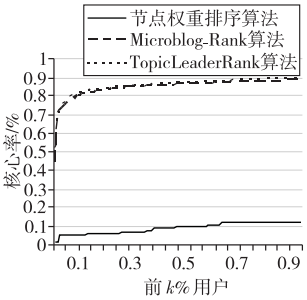

图 6　三种算法在“两会房价”话题下的核心率

覆盖率与核心率仅仅从网络结构和用户交互的角度来评价意见领袖,而忽略了用户本身的属性。因此,本文采用了人工评价的方法对三种算法找出的意见领袖进行综合评价。在此,将排名前 10 的用户当做意见领袖,表 4 列出了三种算法在“单独二胎”话题下得出的意见领袖的相关特征。从表 4 可以看出,节点权重排序算法得出的意见领袖在话题参与度上比较突出,但在其他方面表现一般,特别是在网络特性上表现很差,除了“易富贤”和“阿蚌谈人口”两个用户,其他用户与周围用户的交互都很少,未体现出明显的影响力,同时这些意见领袖在现实中的权威度和知名度都很一般,大部分意见领袖也没有认证。可见,以用户属性为核心的节点权重排序算法更看重在话题中表现活跃的用户,但活跃性仅仅是意见领袖的一个局部

特征,影响力才是意见领袖的主要特征,所以该算法的整体效果比较差。

Microblog-Rank 与 TopicLeaderRank 算法得出的意见领袖在平均质量上都要优于用户权重排序算法,它们找出的意见领袖在各个方面表现都比较优秀。首先这些意见领袖的入度都比较大,表明在相关话题中的影响力比较突出;另外全部意见领袖都通过了认证,大部分意见领袖在真实社会中都具有一定的影响力,例如"人民日报"和"央视新闻";最后这些意见领袖的关注数和粉丝数都比较突出,总微博数也比较多,活跃性比较明显。这说明分析用户交互及其形成的网络结构比单独考虑用户的属性能够更好地挖掘微博话题中的意见领袖。

相对于 Microblog-Rank 算法,TopicLeaderRank 算法得出的意见领袖类型更多样。Microblog-Rank 算法得出的意见领袖除了"协和章蓉娅"之外全是传统社会媒体及网络媒体的微博,而 TopicLeaderRank 算法找出了更多的名人微博,如"张醒生""陆琪"等,同时还找出了有影响力的平台微博"天涯社区"。另外,TopicLeaderRank 算法得出的意见领袖质量比较一致,稳定性很好,而 Microblog-Rank 算法找出的排名第三的意见领袖"法制晚报微调查"并没有比较突出的特征,不属于真正的意见领袖。因此,TopicLeaderRank 算法的整体效果是最好的,同时考虑用户属性、用户交互及其形成的网络结构的算法思路是有效的。

表 4 三种算法在"单独二胎"话题下排名前十的意见领袖信息

算法	排名	昵称	(话题相关)微博数/评论数/转发数	总微博数/关注数/粉丝数/相互关注数	入度	出度	是否认证
节点权重排序算法	1	易富贤	109/58/92	12638/1982/30181/1836	1127	18	是
	2	港丽豪园张涛	61/85/61	127251/2995/9219/2573	26	73	否
	3	多重结构纳米材料	82/126/78	18556/175/547/104	30	42	否
	4	解解明珠	1/195/1	435/191/2187/152	3	187	是
	5	沧海嘉禾	104/54/101	35165/1999/761/190	6	58	否
	6	人口学	49/105/39	3910/883/1000/254	82	109	否
	7	新疆兵团老王	0/174/0	11536/210/642/189	0	1	否
	8	阿蚌谈人口	48/65/38	9886/1780/9446/1543	144	31	是
	9	abc1029039	1/147/0	2303/1125/1024/157	0	76	否
	10	白羊 sysu	93/56/91	8140/170/215/34	2	59	否
Microblog-Rank 算法	1	人民日报	16/4/3	25421/250/17649830/232	19606	1	是
	2	央视新闻	7/1/3	28568/166/15892195/152	6295	1	是
	3	法制晚报微调查	3/0/1	516/64/385/26	18	1	是
	4	财新网	25/0/9	57587/346/1407231/252	2298	2	是
	5	新浪山西	5/2/1	7605/884/576446/229	260	2	是
	6	新浪浙江视频	2/3/1	3636/262/29678/116	1173	3	是
	7	协和章蓉娅	3/3/1	6487/1253/1404917/989	4752	1	是
	8	财新新世纪周刊	8/2/2	11328/889/171643/513	279	1	是
	9	广州日报	12/4/1	50559/928/4405524/506	3880	1	是
	10	人民网	17/0/9	37735/246/11791709/234	4010	1	是
TopicLeaderRank 算法	1	人民日报	16/4/3	25421/250/17649830/232	19606	1	是
	2	头条新闻	22/0/5	85728/237/31197126/211	20042	5	是
	3	央视新闻	7/1/3	28568/166/15892195/152	6295	1	是
	4	南方都市报	16/2/6	34150/381/5825222/329	6797	4	是
	5	协和章蓉娅	3/3/1	6487/1253/1404917/989	4752	1	是
	6	张醒生	2/0/0	17099/658/1507094/570	2753	0	是
	7	天涯社区	7/2/3	7313/252/824611/162	4227	5	是
	8	人民网	17/0/9	37735/246/11791709/234	4010	1	是
	9	陆琪	1/0/0	10917/919/20897119/845	2818	0	是
	10	陈里	1/0/0	15993/1720/20055352/1570	1950	0	是

在表 5 中展示了 TopicLeaderRank 算法在三个不同的话题中挖掘出的意见领袖,即排名前十的用户。可以看出,三个话题的意见领袖基本是不同的,这也说明了在挖掘微博意见领袖时,考虑话题相关性是非常有必要的。

表 5 不同话题下 TopicLeaderRank 算法排名前十的意见领袖

排名	单独二胎	两会雾霾	两会房价
1	人民日报	国家电网	人民日报
2	头条新闻	杨澜	央视新闻
3	央视新闻	中国之声	崔永元
4	南方都市报	南方都市报	摆古论今
5	协和章蓉娅	最高人民法院	新浪江苏
6	张醒生	刘帅 168	烧伤超人阿宝
7	天涯社区	新浪财经	十三點半的 kings
8	人民网	老徐时评	唐僧僧僧僧僧
9	陆琪	挺么么	洗屌少女
10	陈里	华尔街日报中文网	请打草花好吗

4 结束语

本文对微博中的话题意见领袖挖掘进行了研究,并结合用户属性、用户交互和话题特征,提出了一个新的挖掘方法。本文在真实的微博数据上对本文提出的新方法进行实验验证,结果表明该方法是有效的,能够找出典型的话题意见领袖。

在后续的研究中,本文将考虑在算法中进一步融合时间维度和文本内容的话题相关度,从而进一步提高意见领袖挖掘结果的质量。

参考文献:

[1] Li Feng, Du T C. Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs[J]. Decision Support Systems, 2011, 51(1): 190-197.

[2] Grewal R, Mehta R, Kardes F R. The role of the social-identity func-

tion of attitudes in consumer innovativeness and opinion leadership [J]. Journal of Economic Psychology, 2000, 21(3): 233-252.

[3] Gruhl D, Guha R, Liben-Nowell D, *et al.* Information diffusion through blogspace[C]//Proc of the 13th International Conference on World Wide Web. New York: ACM Press, 2004: 491-501.

[4] Zhou Hengmin, Zeng D, Zhang Changli. Finding leaders from opinion networks[C]//Proc of IEEE International Conference on Intelligence and Security Informatics. Piscataway: IEEE Press, 2009: 266-268.

[5] Li Yanyan, Ma Shaoqian, Zhang Yonghe, *et al.* An improved mix framework for opinion leader identification in online learning communities[J]. Knowledge-Based Systems, 2013, 43: 43-51.

[6] Song Xiaodan, Chi Yun, Hino K, *et al.* Identifying opinion leaders in the blogosphere[C]//Proc of the 16th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2007: 971-974.

[7] Miao Qingliang, Zhang Shu, Meng Yao, *et al.* Domain-sensitive opinion leader mining from online review communities[C]//Proc of the 22nd International Conference on World Wide Web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2013: 187-188.

[8] 丁雪峰,胡勇,赵文,等. 网络舆论意见领袖特征研究[J]. 四川大学学报:工程科学版, 2010,42(2): 145-149.

[9] 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(S2): 7-12.

[10] Zhang Weizhe, Li Xiaoqiang, He Hui, *et al.* Identifying network public opinion leaders based on Markov logic networks[J]. The Scientific World Journal, 2014, 2014: 1-8.

[11] Agarwal N, Liu H, Tang L, *et al.* Identifying the influential bloggers in a community[C]//Proc of International Conference on Web Search and Web Data Mining. New York: ACM Press, 2008: 207-218.

[12] 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.

[13] 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(1): 97-106.

[14] Page L, Brin S, Motwani R, *et al.* The pagerank citation ranking: bringing order to the Web[R]. Stanford, California: Stanford InfoLab, 1999.

[15] Yu Xiao, Wei Xu, Lin Xia. Algorithms of BBS opinion leader mining based on sentiment analysis[M]//Web Information Systems and Mining. Berlin: Springer, 2010: 360-369.

[16] Weng Jianshu, Lim E, Jiang Jing, *et al.* TwitterRank: finding topic-sensitive influential Twitterers[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 261-270.

[17] Lin Yan, Li Huaxian, Liu Xueqiao, *et al.* Hot topic propagation model and opinion leader identifying model in microblog network[J]. Abstract and Applied Analysis, 2013, 36(2): 360-367.

[18] 肖宇,许炜,夏霖. 网络社区中的意见领袖特征分析[J]. 计算机工程与科学, 2011, 33(1): 150-156.

[19] 尹衍腾,李学明,蔡孟松. 基于用户关系与属性的微博意见领袖挖掘方法[J]. 计算机工程, 2013, 39(4): 184-189.

[20] 冯时,景珊,杨卓,等. 基于 LDA 模型的中文微博话题意见领袖挖掘[J]. 东北大学学报:自然科学版, 2013, 34(4): 490-494.

(上接第 2663 页)

[3] Liu Chongxin, Liu Tao, Liu Ling, *et al.* A new chaotic attractor[J]. Chaos, Solitons and Fractals, 2004, 22(5): 1031-1038.

[4] 王发强,刘崇新. 分数阶临界混沌系统及电路实验的研究[J]. 物理学报, 2006, 55(8): 3922-3927.

[5] 赵品栋,张晓丹. 一类分数阶混沌系统的研究[J]. 物理学报, 2008, 57(5): 2791-2798.

[6] 刘宗华. 混沌动力学基础及其应用[M]. 北京:高等教育出版社, 2006: 18-50.

[7] Pecora L M, Carroll T L. Synchronization in chaotic systems[J]. Physical Review Letters, 1990, 64(8): 821-825.

[8] Wu Xiangjun, Li Shanzhi. Dynamics analysis and hybrid function projective synchronization of a new chaotic system[J]. Nonlinear Dynamics, 2012, 69(4): 1979-1994.

[9] Čelikovsky S, Chen Guanrong. On the generalized Lorenz canonical form[J]. Chaos, Solitons and Fractals, 2005, 26(5): 1271-1276.

[10] 张合新,范金锁,孟飞,等. 一种新型滑模控制双幂次趋近律[J]. 控制与决策, 2013, 28(2): 289-293.

[11] 张转周,陕振沛,刘衍民. 新三维非线性系统的动力学分析[J]. 应用数学和力学, 2013, 34(12): 1321-1326.

[12] Huang Xiaoling, Ye Guodong. An efficient self-adaptive model for chaotic image encryption algorithm[J]. Communications in Nonlinear Science and Numerical Simulation, 2014, 19(12): 4094-4104.

[13] Wei Zhouchao, Yang Qigui. Dynamical analysis of a new autonomous 3-D chaotic system only with stable equilibria[J]. Nonlinear Analysis, 2011, 12(1): 106-118.

[14] Cai Guoliang, Huang Juanjuan. A new finance chaotic attractor[J]. International Journal of Nonlinear Science, 2007, 3(3): 213-220.

[15] Deng Kuibiao, Li Jing, Yu Simin. Dynamics analysis and synchronization of a new chaotic attractor[J]. Optik, 2014, 125(13): 3071-3075.

[16] Zhou Ping, Huang Kun. A new 4-D non-equilibrium fractional-order chaotic system and its circuit implementation[J]. Communications in Nonlinear Science and Numerical Simulation, 2014, 19(6): 2005-2011.

[17] 尹社会,张勇,徐鹏飞. 一个三维混沌系统的动力学行为及反馈同步[J]. 江西科学, 2013, 31(6): 717-721.

[18] 李国辉,李亚安,杨宏. 基于 EKF 的新混沌系统滤波方法[J]. 系统工程与电子技术, 2013, 35(9): 1830-1835.

[19] 高智中. 一个新自治混沌系统的动力学分析[J]. 数值计算与计算机应用, 2012, 33(1): 1-8.

[20] 孙宇锋,李银. 一个四翼混沌系统的广义控制与同步[J]. 数学的实践与认识, 2014, 44(2): 223-228.