

一种改进拓扑势的意见领袖挖掘算法

毛天铭¹ 关 鹏² 皮德常¹

(南京航空航天大学计算机科学与技术学院 南京 211106)¹ (北京航天飞行控制中心 北京 100094)²

摘 要 意见领袖发现在舆情监测、市场推广、信息传播等领域具有重要的理论指导意义和实际应用价值。针对传统意见领袖挖掘算法片面考虑目标单一属性、缺乏话题相关性、评估缺乏客观性等问题,提出了一种基于改进拓扑势的意见领袖挖掘 ITP 算法。该算法结合具体节点的客观属性和网络结构,采用数据偏差对主观权重进行修正,可客观地对目标节点进行评估,挖掘意见领袖。对真实微博数据集进行实验,结果表明与传统的 3 种方法相比,所提出的算法能挖掘不同背景下的意见领袖,具有较高的相关性和准确度。

关键词 社会网络,意见领袖,权重修正,拓扑势

中图分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.039

Improved Topology-potential-based Opinion Leader Mining Algorithm

MAO Tian-ming¹ GUAN Peng² PI De-chang¹

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)¹

(Beijing Aerospace Control Center, Beijing 100094, China)²

Abstract Opinion leaders mining has important theoretical guiding significance as well as practical application value in many fields such as public opinion monitoring, market promotion and information transmission. As traditional opinion leader mining algorithms have the shortcomings of one-sidedly considering the target property, subjective assessments and lacking of relevance, this paper proposed an improved topology-potential-based opinion leader mining algorithm. Combining with objective attributes and the network structure of specific nodes, this algorithm uses data deviation to correct subjective weight, and evaluates the target node objectively, thereby mining opinion leaders. According to the realistic Microblog statistics, the algorithm proposed in this paper is superior to traditional algorithms in accuracy and relevance, and can mine opinion leaders from different backgrounds.

Keywords Social network, Opinion leader, Weight correction, Topology potential

1 引言

随着互联网和 Web2.0 技术的应用和推广,诸如微博、Twitter、Facebook 等社交服务平台成为人们交流的主要方式之一。这类社交平台很大程度上改变了信息的传播模式,用户通过在社交平台上利用文字、图片、视频等内容进行交互,这一过程不断地产生海量的个人信息和交互数据。从这类社交网站的用户交互数据中挖掘出有效信息逐渐成为数据挖掘领域的热点研究方向。

意见领袖(专家)是指在信息传递和人际互动过程中少数具有较强影响力、活动力的人^[1],他们的言行举止会对相应环境下普通大众产生巨大的影响^[2]。因此,意见领袖的挖掘在舆论监测、市场推广等领域有着重要作用^[3]。现有的意见领袖挖掘相关算法一部分侧重于特定环境下,从话题相关的角度出发,对各目标自身固有属性进行分析;另一部分根据其在网络的重要性进行分析,通过复杂网络的中心性以及 Page-Rank 类算法得出用户的影响力评估。但这两类算法的挖掘

结果都有一定的局限性,不能很好地满足意见领袖挖掘的初衷。本文结合上述两类分析方法,通过对社交网络交互过程中形成的具体话题划分出具体的话题网络,并综合分析话题网络中节点(用户)自身属性以及网络结构属性,得到其价值排序,以此作为依据得出对应话题的意见领袖。

本文第 2 节介绍相关的研究工作;第 3 节对基于改进拓扑势的意见领袖挖掘算法进行了详细的讨论和分析;第 4 节通过具体微博所产生的 3 个话题数据,将本文提出的算法和几种传统的算法进行对比实验分析,验证了算法有效性;最后总结全文并给出进一步的研究方向。

2 相关工作

目前社交网络中意见领袖的挖掘算法主要分为两类:1)基于对研究目标节点自身固有属性的分析;2)侧重从整体网络结构的角度挖掘意见领袖。Duan 等人^[4]通过聚类 and 情感分析的方式来识别股票讨论区中的意见领袖。冯时等人^[5]提出了微博意见领袖衡量模型并整理对应的影响指标,然后采

到稿日期:2015-04-01 返修日期:2015-08-30 本文受国家自然科学基金-民航联合基金(U1433116),航空科学基金(20145752033)资助。

毛天铭(1992—),男,硕士生,主要研究方向为数据挖掘与智能系统,E-mail:1528519021@qq.com;关 鹏(1962—),男,高级工程师,主要研究方向为大数据处理与挖掘技术;皮德常(1971—),男,博士,教授,博士生导师,CCF 会员,主要研究方向为数据挖掘、大数据管理与分析,E-mail:dc_pi@nuaa.edu.cn。

用层次分析法对影响指标进行量化分析,确定各个指标的权重,从而对各节点进行排序,最终得到意见领袖。Twitter^[6]通过用户的粉丝数以及该用户发布消息时被查阅的比例来确定重要节点。基于用户自身属性的意见领袖挖掘算法在评估过程中需要事先给出相应指标的权重因素,这在一定程度上体现决策人员的主观意向,但也缺乏客观性,使得挖掘结果波动程度较大。同时,这类方法普遍缺乏对整个网络系统结构的考虑,存在一定的局限性。基于网络结构的意见领袖挖掘算法以 PageRank 为代表,利用网络拓扑结构进行邻居投票,以此作为评估意见领袖的标准^[7,17]。Weng 等人^[8]通过分析相关话题下用户之间的链接结构来挖掘意见领袖。Haveliwala 等人^[12]对 PageRank 中的评估向量进行线性组合以产生特定话题下的意见领袖。该类算法着重考虑话题网络拓扑结构,无法体现评估人员的主观性,同时缺乏对各节点自身信息的分析。实际经验表明,意见领袖往往是那些能吸引普通大众,且活跃程度较高的节点,所以各节点的出入度、活跃程度等个体信息也有着重要的参考价值。而基于网络结构的挖掘算法往往没有考虑这方面的信息。因此这两类方法都存在一定的局限性和片面性。

目前已有一些通过结合节点固有属性和网络结构来挖掘意见领袖的研究。吴岷辉等人^[3]提出的 OpinionLeaderRank 算法综合考虑了节点固有属性和网络结构,但固有属性权重由决策人员事先设定,使得挖掘结果随权重的不同而波动,缺乏客观性。尹衍腾等人^[9]在挖掘意见领袖的过程中考虑了用户属性以及用户关系,但并未对两类属性综合分析,同时该方法并未对微博的类别进行划分,没有考虑话题相关性。文献^[10-13]中的实验表明根据不同领域划分相应的意见领袖比一般泛化地挖掘意见领袖更具有针对性。Li 等人^[14]提出了一种基于用户属性和局部关系的算法,但缺乏对网络整体结构的考虑。

综合上述方法的优点和不足,本文提出了基于改进拓扑势(Improved Topology-potential-based)的微博意见领袖挖掘算法 ITP,该算法综合考虑了各目标节点自身固有属性与其相应的网络结构,利用从原始数据分布中得出的修正系数来修正主观权重,并得出各节点的个体价值;然后引入拓扑势的概念,并根据实际情况对其进行改进,结合之前得出的个体价值以及相关话题下的网络结构,最终得出各节点的影响力评估值,以此作为微博中意见领袖的判断依据。

3 改进拓扑势的微博意见领袖挖掘算法

本文提出了一种基于改进拓扑势的意见领袖挖掘算法 ITP。ITP 算法主要分为两个阶段:第一阶段根据各目标节点自身固有属性评估各节点的个体价值;第二阶段根据话题网络结构并结合上一步得出的节点个体价值综合得出各节点价值评估,以此获得相关话题的意见领袖。

3.1 节点固有属性分析

在对微博相关话题领域下各节点的评估过程中,只使用单一的固有属性(评论数、好友数、转发数等)来评估意见领袖无疑是片面的;而在基于多属性的综合分析中,相关指标的度量以及权重的设置是影响评价精确度最为重要的两个因素。在实际中,各属性值的度量方式往往不同,为了消除各属性之间不同量纲的影响,对各属性数值进行 min-max 标准化,将

结果值映射到 $[0,1]$ 之间。

$$p^* = \frac{p - p_{\min}}{p_{\max} - p_{\min}} \quad (1)$$

其中, p 为未归一化的原始数据, $p_{\min}$ 为不同节点相关属性的最小值, $p_{\max}$ 为最大值, p^* 为归一化后的属性值。为方便计算,采用如下的评估矩阵 A 来表示不同目标的属性值。

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1,n} \\ a_{21} & a_{22} & \cdots & a_{2,n-1} & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m-1,1} & a_{m-1,2} & \cdots & a_{m-1,n-1} & a_{m-1,n} \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n-1} & a_{m,n} \end{bmatrix}$$

其中, a_{ij} 代表第 i 个节点关于属性 j 的评分($i=1,2,\dots,m$, $j=1,2,\dots,n$)。

得到各节点的属性评估后,需要对各属性值进行加权求和,以得到相应的个体价值。现有的意见领袖挖掘算法大部分采用主观打分的策略来给出相应的权重系数。这类策略能够较好地反映评估人员的主观倾向,但缺点也是明显的,由于缺乏客观性,其评估结果往往波动较大。为考虑评估人员实际需求且保证评估的客观性,先由评估人员给出一组权重系数,再利用属性集之间的离差来修正相应的属性权重。离差反映了各属性在不同节点之间的数值差异,其值越大越能反映该属性的差异程度,相反若离差数值较小,则认为该属性在各节点的区别不大,应减小相应属性的权重。

假设由评估人员给出一组属性集的权重系数 $\bar{w} = \{w_1, w_2, \dots, w_{n-1}, w_n\}$, w_i 代表第 i 个属性的权重。节点 i 与其他节点关于属性 j 的离差设为 d_{ij} 。根据式(2)得到各目标关于属性 j 的离差后,使用式(3)可以得出属性 j 的总离差 D_j ,根据各属性的离差,可以得出各属性的修正系数 r_i ,如式(4)所示。

$$d_{ij} = \sum_{k=1}^n |a_{ij}w_j - a_{kj}w_j| = \sum_{k=1}^n |a_{ij} - a_{kj}|w_j \quad (2)$$

$$D_j = \sum_{i=1}^m d_{ij} = \sum_{i=1}^m \sum_{k=1}^n |a_{ij} - a_{kj}|w_j \quad (3)$$

$$r_i = \frac{n \times D_i}{\sum_{j=1}^n D_j} \quad (4)$$

为方便运算,采用向量的形式来表示修正向量 $\vec{r} = \{r_1, r_2, \dots, r_n\}$ 。根据修正向量,对原权重系数进行修正,得到修正后的权重系数 $\vec{w}'$:

$$\vec{w}' = \{w_1', \dots, w_n'\} = \left\{ \frac{w_1 * r_1}{\vec{w} \cdot \vec{r}}, \dots, \frac{w_n * r_n}{\vec{w} \cdot \vec{r}} \right\} \quad (5)$$

$$value_i = \sum_{j=1}^n a_{ij} \times w_j' \quad (6)$$

分析固有属性的过程中,评估人员可以对不同属性设置相应的权重,以实现评估的主观性;同时为了保证评估的客观性和稳定性,利用原始数据分布修正权重系数,并对各节点属性值进行加权求和得到各节点的个体价值 $value_i$ 。固有属性分析的伪代码如算法 1 所示。

算法 1 固有属性分析

输入:节点属性集 S ,各属性主权重 $\vec{w}$

输出:各节点固有属性评估值集合 V_i

1. for each $s \in S$
2. for each $1 \leq i \leq \text{attribute.size}$
3. if $s_i > \max_i$ then $\max_i = s_i$
4. if $s_i < \min_i$ then $\min_i = s_i$

5. end for
6. end for
7. for each $s \in S$
8. for each $1 \leq i \leq \text{attribute.size}$
9. $s_i = (s_i - \min_i) / (\max_i - \min_i)$
10. end for
11. end for
12. for each $1 \leq i \leq \text{attribute.size}$
13. 按式(2)计算属性离差 dev_i
14. end for
15. for each $1 \leq i \leq \text{attribute.size}$
16. $r_i = (\text{attribute.size} * \text{dev}_i) / \sum_{j=1}^{\text{attribute.size}} \text{dev}_j$
17. end for
18. for each $w \in W$
19. $w_i' = (w_i * r_i) / \sum_{j=1}^{w.\text{length}} w_j * r_j$
20. end for
21. for each $s \in S$
22. $v_i = \sum_{j=1}^{s.\text{size}} s_j * w_j'$
23. end for
24. return V

在分析节点属性的过程中,算法首先对原始数据进行归一化预处理,然后再利用离差对主观权重进行修正,从而弥补了基于节点自身属性的意见领袖挖掘算法缺乏客观性的缺点,但在分析过程中,并未考虑话题的网络结构特征,因此在下一节结合节点自身属性和网络结构,对其进行综合评估分析。

3.2 节点综合属性分析

在分析相关话题各节点的网络属性前,需要从实际的微博信息交互中构建相应的话题网络 $G=(V, E)$,其中 V 代表节点(用户)集, E 代表节点(用户)之间的有向边集,网络关系如图 1 所示。

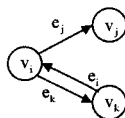


图 1 网络关系图

节点 v_i 分别有一条指向节点 v_j 和节点 v_k 的有向边 e_j , e_k , 节点 v_k 有一条指向节点 v_i 的有向边 e_i , 边的权重由实际联系强度确定。网络构建完成后,需要对该话题网络中各节点进行综合属性的分析和评估。为客观描述网络中各节点的重要程度,引入拓扑势的概念^[15]。根据数据场的势函数定义,任意节点的拓扑势可表示为:

$$\phi(v_i) = \sum_{j=1}^n (m_j \times e^{-\frac{d_{ij}^2}{\sigma}}) \quad (7)$$

其中, m_j 为标号为 j 的节点的质量, e 为自然常数, d_{ij} 为节点 i 到节点 j 的距离, σ 为影响因子。

拓扑势描述的是势场中节点对其他节点的影响能力,这与我们进行意见领袖挖掘的初衷是一致的。拓扑势的引入为我们提供了一种客观的度量方法,但物理势场和实际的微博话题网络有一定的差异。由上述定义可知,传统拓扑势在计算的过程中只考虑了其他节点的“质量”,而实际情况中,节点固有属性的差异也需要我们认真考虑。以微博为例,在网络结构相同的情况下,拥有较大个体价值的节点往往比那些个

体价值较低的节点拥有更多的话语权,即那些拥有更多粉丝数、更多被收藏次数的节点往往是潜在的意见领袖。因此,需要对传统的拓扑势进行改进,即在计算过程中考虑目标节点自身的属性。改进后的拓扑势如式(8)所示。

$$\phi(v_i) = \sum_{j=1}^n (m_i \times m_j \times e^{-\frac{d_{ij}^2}{\sigma}}) = m_i \times \sum_{j=1}^n (m_j \times e^{-\frac{d_{ij}^2}{\sigma}}) \quad (8)$$

将属性分析中所得个体价值视为节点的“质量”,对由实际话题交互信息构建的图模型进行处理得到各节点之间的最短路径,最终得到各节点的价值评估,给出相关话题的意见领袖。综合分析过程的伪代码如算法 2 所示。

算法 2 综合分析

输入: 节点个体价值评估集 V , 用户关联记录集 R

输出: 话题相关意见领袖集 A

1. $E = \emptyset, T = \emptyset$
2. for each $r \in R$
3. $E = E \cup \{r\}$
4. end for
5. for each $1 \leq i \leq U.\text{size}$
6. $D = \text{Dijkstra}(E, u)$
7. 使用式(8)计算 t_i
8. $T = T \cup \{t_i\}$
9. end for
10. sort(T)
11. $A = \text{getTopK}(T)$
12. return A

上述算法首先根据用户关联集构造相应的话题网络,然后计算各节点之间的最短距离,并根据式(8)评估各节点的综合价值,最后将节点的评估集进行排序,取有序集中前 k 位作为推荐的意见领袖。

综合上述分析,基于改进拓扑势的意见领袖挖掘算法在评估的过程中,综合分析了节点的固有属性并且根据数据离差修正了主观权重,同时考虑了相关话题的网络结构,使得挖掘结果能与意见领袖的定义相符。

4 实验结果与分析

本节将本文提出的 ITP 算法与几种具有代表性的意见领袖挖掘算法进行分析和比较,其中包括传统复杂网络理论中基于节点出入度 IODegree 的意见领袖挖掘算法^[16]、根据评估人员给出的权重系数对各节点属性加权求和的 AHP 算法^[1],以及基于网络结构的邻居投票 PageRank 算法^[17]。将 ITP 算法两个阶段的意见领袖候选集与传统算法的结果集进行对比分析,以验证 ITP 算法的有效性。

4.1 实验数据

本文实验数据来源于新浪微博 2014 年 5 月 3 个热点话题的真实微博信息,其中每个话题包括相应的用户信息、用户关系信息以及微博信息。详细信息统计如表 1 所列。

表 1 原始话题数据集信息

话题	参与用户数	微博数	用户关联数
公务员	7673	9048	60168
雾霾	6157	7183	51782
同桌的你	10315	10886	53790

为了验证 ITP 算法的有效性,将 ITP 所得实验结果与只

考虑节点自身属性的 IODegree 算法、AHP 层次分析法以及考虑网络结构的 PageRank 算法进行对比与分析。属性分析中涉及到的实际用户属性有微博转发数、微博评论数、微博被赞数,相关话题属性统计信息如表 2 所列。

表 2 话题属性统计信息

话题	微博转发数	微博评论数	微博被赞数
	Max/Min/Dev	Max/Min/Dev	Max/Min/Dev
公务员	1188415/0/13365.6	182041/0/11097.2	70466/0/20922.2
雾霾	123273/0/50528.2	36686/0/27184.9	142968/0/12161.6
同桌的你	271087/0/56923.0	133402/0/25399.8	206206/0/25747.0

对原始数据进行筛选和预处理后,将各属性权重设置为属性个数的倒数。将基于节点自身属性分析的结果与综合分析的结果进行对比,如表 3 所示。

通过表 3 可以发现,基于节点自身属性的分析结果中,大部分节点属于公众明星类,这类节点因为自身的特点,有着数量众多的关注者,因此该类节点在基于属性的分析中往往有着较高的排名。但这类节点发布的微博信息往往并不具有话

题相关性,也不能在信息传播的过程中起到主导作用,这也是基于属性分析挖掘算法的普遍缺点。通过对比可以发现,在“公务员”、“雾霾”这类大众相关的话题背景下,ITP 算法得到的综合分析结果中,类似“人民日报”、“央视新闻”这类受众广泛、对普通大众具有导向作用的节点排名占据了前几位,而诸如“王俊凯”、“胡歌”等娱乐明星类节点排名大幅降低,同时“陈里”、“糖不吃先生”这类话题相关的节点排名也得到了较大的提高,这与实际情况是一致的。而对“同桌的你”这一娱乐话题的属性分析和综合分析的结果基本一致,这是因为该话题与娱乐明星类节点高度相关,该类节点往往引导着该类话题的传播。同时,电影主演“林更新”、“周冬雨”、制片人“刘同”这类节点排名均有一定的上升,提高了候选节点的话题相关性,使排序结果更具有说服力。

为了验证 ITP 算法的有效性,将其与传统的 3 种意见领袖挖掘算法进行对比。“公务员”话题下评估价值排名前 10 的节点如表 4 所列。

表 3 实验结果对比

排名	公务员		雾霾		同桌的你	
	固有属性分析	综合分析	固有属性分析	综合分析	固有属性分析	综合分析
1	锦鲤大王	人民日报	易烊千玺	央视新闻	郭敬明	林更新
2	王俊凯	头条新闻	央视新闻	人民日报	GEM 邓紫棋	韩寒
3	唐嫣	央视新闻	家乐福中国	当时我就震惊了	林更新	GEM 邓紫棋
4	人民日报	GEM 邓紫棋	互联网的那点事	何炅	新浪娱乐	刘同
5	GEM 邓紫棋	新浪娱乐	李晨	互联网的那点事	刘同	周冬雨
6	胡歌	财经网	武汉日记	倒吸一口凉气君	思想聚焦	新浪娱乐
7	头条新闻	陈里	人民日报	财经网	周冬雨	郭敬明
8	央视新闻	胡歌	倒吸一口凉气君	糖不吃先生	韩寒	思想聚焦
9	边伯贤私人微博	新浪搞笑排行榜	玩转 instagram	新浪娱乐	马伊琍	同桌的你电影
10	中流青年	新闻晨报	当时我就震惊了	头条新闻	同桌的你电影	魏晨

表 4 “公务员”话题意见领袖

排名	IODegree	AHP	PageRank	ITP
1	头条新闻	锦鲤大王	媒体微博助理	人民日报
2	韩寒	人民日报	央视新闻	头条新闻
3	央视新闻	TFBOYS	人民日报	央视新闻
4	人民日报	头条新闻	央视财经	GEM 邓紫棋
5	梦想家林志颖	运筹帷幄唯你	公安部打四黑除四害	新浪娱乐
6	留几手	中公教育图书	环球时报	财经网
7	GEM 邓紫棋	财经网	中国之声	陈里
8	何炅	山东华图	新华视点	胡歌
9	张亮 Sean	考德上教育	新华社中国网事	新浪搞笑排行榜
10	李开复	广东华图	网络新闻联播	新闻晨报

同时对“雾霾”和“同桌的你”这两个微博话题进行对比分析,分别取各算法评估结果中排名前 5% 的节点作为候选的意见领袖,并统计不同类别节点所占比例。以“雾霾”话题下 ITP 算法所得的意见领袖候选集为例:前 5% 节点 308 例,其中官方平台类节点 78 例,公众人物类节点 96 例,话题相关类节点 71 例,娱乐信息类节点 47 例,其他节点 16 例。将上述数值除以候选集数 308,即得各类别节点所占比例。ITP、AHP 等 4 类算法在相关话题背景下,所得候选集各类节点所占比例如图 2 和图 3 所示。

从表 4 以及图 2、图 3 的实验结果可以发现 IODegree 算法给出的意见领袖偏向于那些固有属性值较高的节点。“头条新闻”、“央视新闻”、“人民日报”等官方平台账号有较高的

排名,但挖掘结果中更多的是娱乐相关的明星类人物,这类用户由于自身的特性往往会吸引很多普通大众的关注,使其有着较高的排名,但这类节点的微博内容与主题的相关性不大,缺乏说服力。AHP 层次分析法的结果有着明显的话题偏向性,从表 4 可以发现,其结果中除了有“人民日报”、“头条新闻”这类与“公务员”话题密切相关的官方公共类节点外,也有 5 个与公务员话题相关的教育图书类节点,这类节点发布的信息中,很多内容包含“公务员”这一关键字,因此在微博信息发送量、话题持续关注度上有明显的优势,这也导致其排名较高。但实际中,这类节点并不会得到普通大众的关注,其影响范围缺乏普适性,与我们意见领袖挖掘的初衷不符。从 PageRank 算法的结果中可以发现候选意见领袖基本为官方性质的微博,这类节点因为自身的特点有着不低的关注度,且相互联系较为紧密,使得基于网络结构的算法结果侧重此类节点,这与 PageRank 的投票思想是相符的;但缺点也是显而易见的,由于微博内容的单一和相似,且内容单调、缺乏话题相关性,不能长期保持对普通大众的吸引。从以上分析以及图 2、图 3 可以发现,传统的 3 种方法因为意见领袖挖掘的侧重点不同,使得结果有失偏颇,即挖掘结果偏向特定的一类节点,这与我们的初衷相违背。ITP 算法同时考虑了节点固有属性和网络结构两类因素,例如“人民日报”、“头条新闻”等这些官方微博占有部分比重,且诸如“新浪娱乐”、“财经网”这种普通公众关注度较高的娱乐信息类微博也有相应的体现。同时值得注意的是,“陈里”这类在普通大众关注度不高的节点,因为

和特定话题的高度相关,在该话题下也有着较高的排名,这也体现了 ITP 算法挖掘结果的话题相关性。综上所述,ITP 算法通过综合分析相关话题下各目标节点的个体属性和网络结构,弥补了传统方法挖掘结果单一、缺乏话题相关性的缺点,使得结果符合意见领袖挖掘的初衷,证明了该算法的有效性。

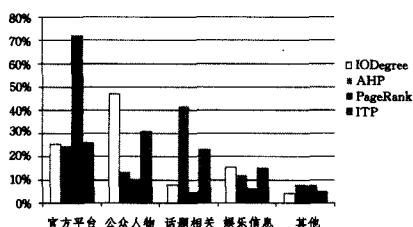


图2 “雾霾”话题下分析结果

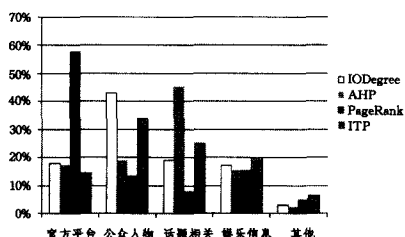


图3 “同桌的你”话题下分析结果

结束语 本文研究了微博中话题相关的意见领袖挖掘问题,针对现有方法依据单一固有属性或网络属性进行分析的局限性,提出了一种基于拓扑势的意见领袖挖掘算法,同时在评估过程中,利用实际数据对主观权重进行适当修正,保证评估的客观性。通过实际数据和对比实验表明,本文提出的 ITP 算法能较好地处理不同话题背景下的意见领袖挖掘,提高了结果的相关性和准确度。在后续的工作中,将进一步研究权重系数修正和网络图中有向边权重的评估,以获得更好的、符合实际的挖掘效果。

参考文献

- [1] Liu Zhi-ming, Liu Lu. Recognition and Analysis of Opinion Leaders in Microblog Public Opinions[J]. Systems Engineering, 2011, 29(6): 8-16 (in Chinese)
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
- [2] Zhao Yu, Xu Wei, Xia Lin. A Feature Analysis of the Opinion Leader in On-Line Communities [J]. Computer Engineering & Science, 2011, 33(1): 150-156 (in Chinese)
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
- [3] Wu Xian-hui, Zhang Hui, Yang Chun-ming, et al. An Algorithm of Topic-related Microblogging Opinion Leaders Mining [J]. Journal of Chinese Computer Systems, 2014, 38(10): 2296-2301 (in Chinese)
吴岷辉,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, 38(10): 2296-2301
- [4] Duan Jiang-jiao, Zeng Jian-ping, Luo Bang-hui. Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis[C]// Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). Warsaw, Poland, 2014: 377-383
- [5] Feng Shi, Jing Shan, Yang Zhuo, et al. Detecting topical opinion leaders based on LDA model in chinese microblogs[J]. Journal of Northeastern University(Natural Science), 2013, 34(4): 490-494 (in Chinese)
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
- [6] Zhang Wei-zhe, He Hui, Cao Bo-ran. Identifying and evaluating the internet opinion leader community based on k-clique clustering[J]. Neural Computing and Applications, 2014, 25 (3-4): 595-602
- [7] Lin Yan, Li Hua-xian, Liu Xue-qiao, et al. Hot topic propagation model and opinion leader identifying model in microblog network [J]. Abstract and Applied Analysis, 2013(6): 1-13
- [8] Weng Jian-shu, Lim E, Jiang Jing, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. 2010: 261-270
- [9] Yin Yan-teng, Li Xue-ming, Cai Meng-song. Mining method of microblog opinion leader based on user relationship and attribute[J]. Computer Engineering, 2013, 39(4): 184-189 (in Chinese)
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
- [10] Miao Qing-liang, Zhang Shu, Meng Yao, et al. Domain-sensitive opinion leader mining from online review communities[C]// Proceedings of the 22nd International Conference on World Wide Web. 2013: 187-188
- [11] Bi Bin, Tian Yuan-yu, Sismanis Y, et al. Scalable Topic-Specific Influence Analysis on Microblogs[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. 2014: 513-522
- [12] Haveliwala T H. Topic-Sensitive PageRank[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784-796
- [13] Bi Bin, Kao Ben, Wan Chang, et al. Who are experts specializing in landscape photography? Analyzing Topic-specific Authority on Content Sharing Services[C]// Proceedings of the The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 1506-1515
- [14] 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
- [15] Gan Wen-yan, He Nan, Li De-yi, et al. Community Discovery Method in Networks Based on Topological Potential[J]. Journal of Software, 2009, 20(8): 2241-2254 (in Chinese)
淦文燕,赫南,李德毅,等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254
- [16] Zhang J, Ackerman M S, Adamic L. Expertise networks in online communities: structure and algorithms[C]// Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 221-230
- [17] Zhou H, Zeng D, Zhang C. Finding leaders from opinion networks[C]// IEEE International Conference on Intelligence and Security Informatics, 2009 (ISI'09). IEEE, 2009: 266-268