基于多维特征分析的社交网络意见领袖挖掘

曹玖新,陈高君,吴江林,刘 波,周 涛,胥 帅,朱子青
(1. 东南大学计算机科学与工程学院,江苏南京 210096; 2. 东南大学计算机网络和信息集成教育部重点实验室,江苏南京 210096)

摘 要: 在社交网络中进行意见领袖的挖掘对信息传播与演化的深度分析、舆情监控和引导具有重要意义,本文综合结构特征、行为特征和用户的情感特征对意见领袖节点挖掘问题进行研究. 本文首先对微博真实文本数据进行话题识别得到主题社区,在主题社区中基于用户节点之间的关注关系构建交互网络拓扑. 然后分别从结构、行为和情感三个维度对用户的影响力进行度量. 最后,分析用户在主题社区中的影响力分布与传播规律,提出意见领袖识别算法 MFP(Multi-Feature PageRank). 实验表明,该算法可有效地挖掘潜在的意见领袖节点,能够获得较高的支持率.

关键词: 社交网络; 话题; 情感分析; 意见领袖
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2016) 04-0898-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.04.021

Multi-Feature Based Opinion Leader Mining in Social Networks

CAO Jiu-xin, CHEN Gao-jun, WU Jiang-lin, LIU Bo, ZHOU Tao, XU Shuai, ZHU Zi-qing
(1. *School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu* 210096, *China*; 2. *Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing, Jiangsu* 210096, *China*)

Abstract: Mining opinion leaders in social network is important for analysis of information dissemination and evolution of public opinion. This paper conducts the study on this problem considering structural features, behavior and emotional characteristics comprehensively. Firstly, we extract topics from micro-blogging texts, and get user communities according to the topic division, and an interactive network topology of topic community is built with the following relationships. Then, three kinds of user feature are gained from different aspect: network structure, user behavior and user sentiment. Finally, according to the analysis of users' influence distribution, opinion leaders mining algorithm MFP (Multi-Feature PageRank) is proposed. Experiments show that the algorithm can obtain the potential opinion leader nodes effectively, and have a good performance in support rate from other user nodes.

Key words: social network; topic; sentiment analysis; opinion leader

1 引言

社交网络是由社会成员之间的相互交互所形成的相对稳定的社会结构[1],具有复杂的网络结构和信息动态传播机制. 随着互联网的普及和发展,在线社交网络已经成为人们结识好友和共享信息的主要渠道,并已演变为一个巨大的社会热点事件发布平台. 社交网络节点的异质性决定了节点地位的非对等性,部分节点对其他节点具有较大的影响力,对舆论的发展能起到关键性的导向作用,被称为"意见领袖"(Opinion Leader). 在社交网络中,用户发表的言论往往受到一段时期内直接相关的事件或活动影响,与特定主题(Topic)紧密相关. 因此,社交网络中的意见领袖挖掘是面向主题社区的. 在主题社区"意见领袖"的影响下,热点新闻或热点信息会吸引众多的用户参与讨论,产生大量反馈、交互和评价,形成热点话题. 各种观点逐渐被引导聚合,形成具有某些倾向性的结果. 因此,对意见领袖的挖掘研究,有助于社交网络中的信息传播与演化的深度分析,为社交网络的舆情监控和引导提供决策依据

收稿日期: 2014-10-16; 修回日期: 2015-06-12; 责任编辑: 李勇锋
基金项目: 国家863高技术研究发展计划(No. 2013AA013503); 东南大学计算机网络和信息集成教育部重点实验室基金(No. 93K-9); 国家973重点基础研究发展计划(No. 2010CB328104); 国家自然科学基金(No. 61272531, No. 61202449, No. 61272054, No. 61370207, No. 61370208, No. 61300024, No. 61320106007, No. 61472081); 高等学校博士点学科专项科研基金(No. 20110092130002); 江苏省科技计划基金(No. SBY2014021039-10); 江苏省网络与信息安全重点实验室基金(No. BM2003201)

和技术支撑.

2　相关工作

关于社交网络中意见领袖的挖掘,研究者重点关注图结构、用户内容、用户行为记录等多个方面,综合运用了社会网络理论和各类机器学习方法. 研究对象既涵盖了传统 BBS 网络、博客网络,也包括 Weibo、Twitter 等微博类网络. 近来,在线股票平台、在线学习平台等具有用户交互的专业性平台由于其社交网络属性,也成为了研究热点.

现有关于意见领袖挖掘的研究侧重点各有不同. 文献[2~4]基于社交网络结构,利用节点入度、中介中心性、接近中心性等特征来衡量节点影响力,但是其准确度不高,不能很好地体现节点的影响力. 文献[5~7]通过构建社交关系网络并基于用户行为和兴趣领域发现社区中的意见领袖. 文献[8~13]从用户发表的内容出发,分析文本语义信息,挖掘用户潜在情感,进而找到社区中的意见领袖. 在微博社交网络中,综合考虑网络拓扑结构、话题语义和文本情感因素,对研究新的节点特征模型与设计意见领袖挖掘算法具有重要意义.

3　意见领袖挖掘框架

本文将微博社交网络中节点之间的关系分为物理关系和虚拟关系,物理关系如关注关系、回复关系等,虚拟关系如兴趣相似关系和观点相似关系. 基于对以上两种关系的分析,本文利用话题识别、情感分析等技术,在微博社交网络中分析用户节点的影响力,挖掘意见领袖节点.

本文总体技术框架如图 1 所示,主要包括以下三个方面的工作:

第一,主题社区发现. 通过对用户产生的微博进行文本分析,采用聚类算法进行话题识别,得到话题集后进行话题-用户映射,形成主题社区.

第二,意见领袖节点特征分析. 不仅分析用户节点的结构特征和行为特征,而且对微博文本的语义情感进行分析,得到用户之间的情感极性.

第三,意见领袖节点识别算法. 在分析用户节点特征的基础上,设计意见领袖识别算法.

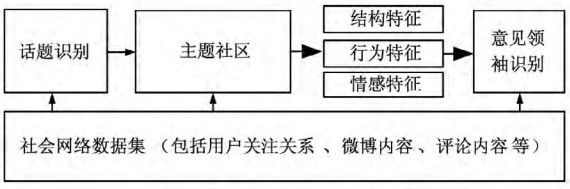

图1　意见领袖挖掘总体框架

4　主题社区发现

在微博社交网络中,用户所发表、转发和回复的微博都与特定的话题相关,表现了用户的兴趣爱好. 具有相似兴趣爱好的用户往往会形成一个以"兴趣话题"为核心的虚拟社区,本文称之为主题社区. 主题社区中的话题特征为用户结点的特征提取提供了丰富的语义支持,有助于意见领袖挖掘算法的设计.

4.1　话题识别

话题识别是主题社区发现的前提,是通过真实的社交网络数据进行分析找出热点话题的过程. 本文采用无监督且自适应的话题识别方法,其思想是根据话题识别过程中的相关反馈对话题模型进行自学习,为话题模型引入新的特征,同时调整特征权重,减少先验知识的稀疏性对话题模型的影响. 话题识别的具体流程如算法 1 所示:

算法 1　话题识别算法

输入: 微博集 $W=\{w_0,w_1,w_2\cdots w_n\}$

输出: 话题集 $T=\{t_0,t_1,t_2\cdots t_n\}$

话题微博集: $C_0,C_1,C_2\cdots C_n$

1　将微博集中的微博都处理成文本向量 v_i,并构成文本向量集 V

2　初始化话题集为 $T=\{t_0\}$,其中 $t_0=v_0$

3　对于文本集 V 中的每一条文本 v_i,计算其与已有话题相似度:

$$s_{ij}=similarity(v_i,t_j)$$

4　如果 $s_{i\max}>\varepsilon$,则更新对应话题下的微博,否则定义新的话题 t_{new} 并且更新话题集 $T=\{t_{new}\}\cup T$

微博文本向量 $v_i=(tfidf_0,tfidf_1,tfidf_2\cdots ifidf_n)$,其中 $tfidf_i$ 表示分词 i 在文本语料库中的词频-逆文档频率值(Term Frequency-Inverse Document Frequency,TF-IDF). 计算方法如式(1):

$$tfidf_{i,j}=tf_{i,j}*\ idf_i*\ len_i=\frac{n_{i,j}}{\sum_{k=1}^{K}n_{k.j}}*\log\frac{|C|}{|\{c:w_i\in c\}|}*\ len_i \quad (1)$$

其中 $n_{i,j}$ 表示分词 i 在文档 j 中的出现的频次,$\sum_{k=1}^{K}n_{k.j}$ 为文档 j 中所有分词的频次总和,K 为文档 j 中的分词总数,$|C|$ 表示文档总数,$|\{c:w_i\in c\}|$ 表示包含该分词 i 的文档数量,len_i 是分词 i 的长度.

4.2　主题社区网络拓扑构建

基于话题识别的结果,根据与特定话题 t_i 相关的微博集合 C_i 可构建相应的交互网络拓扑,其构建过程可用图 2 形象地描述. 首先,在 B 层中找出 C_i 中所有微博的作者,形成一个由用户节点组成的集合 U. 然后,抽取原始数据集中这些用户的交互关系,添加到 U 中,得到主题社区交互网络拓扑,如图 2 中 C 层.

网络拓扑构建算法如算法 2 所示.

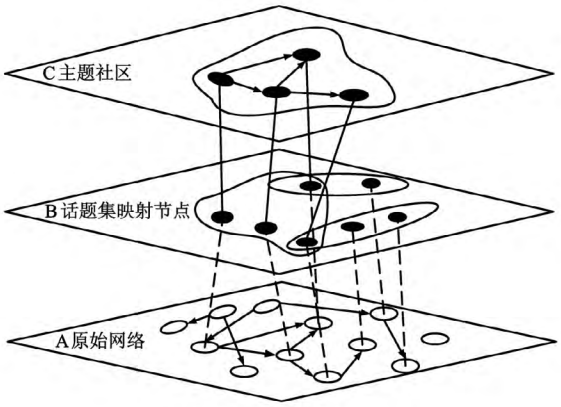

图2 主题社区构建过程

算法2 网络拓扑构建算法

输入: 话题微博集对应的用户种子集合 U

输出: 网络拓扑(节点集合 V、边关系集合 E、边权重)

```
将话题微博集对应的用户种子集合 U 依次放入队列 Q
while( 队列 Q 非空且用户数未满足要求)
do
    从队列 Q 首部取出用户节点 u_i;
    抽取出该用户节点 u_i 所发表的属于这个话题集的微博集合 M;
    对 M 中每一条微博 mi:
        获取评论微博 m_i 且是 u_i 好友的用户集合 US
        ( 若采用转发关系,则获取转发微博 m_i 的下一跳用户集合 US)
        对 US 中每个用户 u_j:
            若(u_j,u_i) ∈ E,则 W_{j,i} = W_{j,i} + 1
            否则建立新边(u_j,u_i) 并令 W_{j,i} = 1
        将 u_j 加入节点集合 V 以及队列 Q
end while
```

5 意见领袖节点特征分析

主题社区构建之后,挖掘用户影响力的特征成为意见领袖节点识别的关键因素. 用户节点的影响力是多种复杂因素共同作用的结果. 基于对用户节点的深度分析,综合用户节点的各类属性,本文选取用户的结构特征、行为特征和情感特征作为用户影响力特征.

5.1 结构特征

结构特征体现了用户本身因素和所在网络拓扑的结构因素,如用户的好友数、粉丝数、中介中心度. 根据社交网络拓扑模型可以得出特征量值,并作归一化处理,这里采用最大最小值归一化方法. 假设特征值量化后为 f,最大值为 $f_{\max}$,最小值为 $f_{\min}$,则归一化后的 f_n 为:

$$f_n = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{2}$$

取归一化后的数值的平均值,作为用户的结构特征值:

$$S(u) = (u_{\text{friend}} + u_{\text{follower}} + u_{\text{betweeness}})/3 \tag{3}$$

其中 u_{friend} 是好友数归一化后数值,u_{follower} 是粉丝数归一化后数值,$u_{\text{betweeness}}$ 是中介中心度归一化后的数值.

5.2 行为特征

行为特征归结为以下两点:

(1) 活跃度: 用户在一段时间内主动发表、转发、评论的有效微博条数;

(2) 传播力: 用户的微博被转发、被评论的有效条数.

在网络拓扑结构中,将用户的活跃度和传播力这两个特征量化并归一化后取其平均值得到用户的行为特征值.

$$B(u) = (u_{\text{active}} + u_{\text{forward}})/2 \tag{4}$$

其中 u_{active} 是用户主动发表、转发、评论的有效微博条数归一化后的数值,表示用户活跃度,u_{forward} 是用户的微博被转发、被评论的有效条数归一化后的数值,表示用户传播力.

5.3 情感特征

在意见领袖挖掘问题中,用户的影响力不能简单地从结构特征和行为特征衡量,还需要从语义内容角度去评价特定用户对于某一话题的观点[14],这便是用户的情感特征. 在微博社交网络中,原创微博会引发大量的评论微博,这些评论综合体现了众人的褒贬情感,因此本文将评论微博作为情感特征分析的对象.

本文采用基于情感词典的机器学习方法对微博评论进行情感极性分析,并将情感极性分为正向极性、中性和负向极性三类. 常见的情感词典有台湾大学中文情感词典(NTUSD)[15]、知网(HowNet)[16],哈尔滨工业大学《同义词词林扩展版》[17]等,本文选用知网词典.

本文采用的微博评论情感极性挖掘的特征如表1所示. 使用经过标记的250000条正向情感微博和250000条负向情感微博并基于决策树方法训练分类器,采取十次十折交叉验证的方法保证分类器的性能.

根据决策树分类模型,对主题社区的原创微博的评论进行情感极性分析,获得用户之间的意见趋势,具体的流程如图3所示.

获得评论的情感极性分类后,设用户 u_i 关注了用户 u_j,u_i 多次评论了 u_j 所发表的微博,将 u_i 对 u_j 的正向评论占总评论的比例定义为情感支持权重:

$$W_{ij} = \frac{N_{\text{pos}}}{N_{\text{total}}} \tag{5}$$

其中,N_{pos} 表示 u_i 对 u_j 进行正向评价的次数,N_{total} 表示 u_i 对 u_j 的总评价数.

表 1　用于情感分析的特征

	特征	描述
1	情感词	正向情感词个数
2		负向情感词个数
3	否定词	文本中出现否定词个数
4	程度副词	文本中出现程度副词个数
5	带有情感色彩的标点符号个数	文本中出现“!”或者“?”个数
6	表情符号	包含正向表情符号个数
7		包含负向表情符号个数

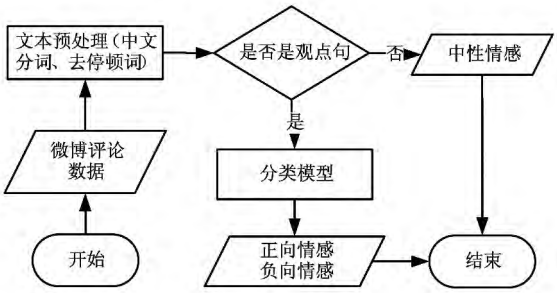

图3　微博文本情感极性分析

6　意见领袖挖掘算法

本文在主题社区中挖掘意见领袖节点,综合考虑用户节点的结构特征、行为特征和情感特征,提出多维意见领袖挖掘算法 MFP(Multi-Feature PageRank),算法使用式(6)计算节点的意见影响力值:

$$INF(u_i) = (1-d)(S(u_i) + B(u_i)) + d^* \sum_{j \in Followers(i)} \frac{W_{ji}}{\sum_{k \in Friends(j)} W_{jk}} INF(u_j) \quad (6)$$

其中 $INF(u_i)$ 为用户的影响力值,$S(u_i)$ 为用户归一化后的结构特征,$B(u_i)$ 为用户归一化后的行为特征,$W_{j,i}$ 为 u_j 对 u_i 的意见权重,d 为阻尼系数,取 0.8.

MFP 算法的提出借鉴了网页重要性排名算法 PageRank 的思想,认为一个用户的意见影响力不仅与其结构特征和行为特征紧密相关,还取决于其粉丝用户的意见态度:如果粉丝用户对该用户的意见普遍赞成,则相应的意见权重也就越大,对于此用户的影响力贡献度越大.另一方面,当前用户的意见影响力大小还与他的邻居用户的意见影响力有关,如果邻居用户的意见影响力普遍较高,而且对当前用户表现出正向情感,则对当前用户的意见影响力会有很大贡献.因此,MFP 算法既具有 PageRank 的优势,又结合语义内容,能发现深层次的影响因素.算法具体描述如下:

算法 3　多维意见领袖挖掘算法(MFP)
输入:网络拓扑关系
　　(节点集合 V 和边关系集合 E 及情感权重 W)
输出:节点影响力排名(Top-K)
1　计算 v_i 的邻居节点的边情感权重之和 $Self(v_i)$:
　　$Self(v_i) = (1-d)(S(v_i) + B(v_i))$
2　初始化 ε,令 $\varepsilon > \varepsilon_{threshold}$
3　while($\varepsilon > \varepsilon_{threshold}$)
4　　do
5　　$\varepsilon = 0$
6　　计算粉丝节点 v_j 对 v_i 的影响力贡献:
　　$INF(v_i) = INF_{old}(v_i) + d^* \frac{W_{ji}}{Self(v_j)}^* INF_{old}(v_j)$
7　　计算前后两次迭代的差值之和:
　　$\varepsilon += |INF(v_i) - INF_{old}(v_i)|$
8　end while
9　输出影响力较大的 K 个节点

假设拓扑模型中的用户总数为 N,迭代次数为 k,则上述 MFP 算法的时间复杂度为 $O(kN^2)$.

7　实验

7.1　实验数据

本文基于新浪微博的开放 API 抓取实验数据,抓取程序依照广度优先的策略,从一个特定的用户开始,爬取该用户最近发表的 100 条微博,对于其中的每条微博,再爬取该微博的转发微博以及转发该微博的用户信息,将这些用户添加至待爬取队列.结束对一个用户的处理之后,取出待爬取用户队列的第一个用户,继续同样的处理,循环往复.最终获取的数据集中共包括 10785921 条微博,其中 28.98% 是原创微博.

7.2　主题社区发现及特征值计算

微博文本大部分内容简短、偏口语化,并夹杂表情符号,这种文本特点会导致话题识别的准确度不高.为此,需要对微博数据集做如下处理:

(1)去除微博中表示表情的词;
(2)去除停用词;
(3)排除长度小于 100 的微博.

根据上述处理方法,在 10785921 条微博中共得到 792051 条微博,使用算法 1 对 792051 条微博进行话题识别,得到了 1276 个话题.其中规模最大的话题为“北京暴雨”,包含 877 条微博.

根据网络拓扑构建算法,这里选取最大规模话题集“北京暴雨”中的用户节点为初始种子节点,分别使用用户之间的关注关系和转发关系,进行网络拓扑的构建,表 2 是关注关系网络和转发关系网络的各项网络指标.

表 2 两种关系网络的指标对比

	关注关系网络	转发关系网络
节点数	7865	105024
边数	225830	115269
平均度	28.713	1.098
网络直径	11	34
平均路径长度	3.503	11.594
平均聚类系数	0.246	0.006

从表 2 可以看出,转发关系网络的平均度很低,网络直径很大,造成转发关系网络的紧密程度不高. 而关注关系网络用户之间的关系更加紧密,符合标准社交网络的一般规律. 基于意见领袖挖掘研究的特点,本文采用关注关系网络作为意见领袖挖掘的基本图模型,其较小的网络直径和较大的聚类系数更加符合主题社区的构建需求.

使用本文第 5 节中的情感分析模型对评论微博进行极性分类,统计出主题社区内所有用户的情感支持权重. 图 4 为用户情感支持权重的分布情况.

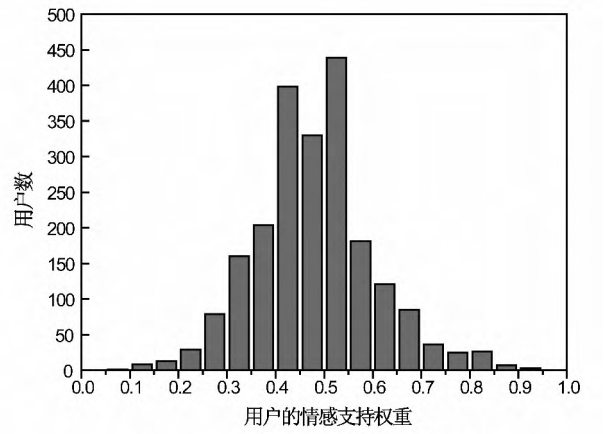

图4 主题社区中用户的情感支持权重分布

由上图可以看出,情感支持权重在 0.4 ~ 0.55 范围之间的用户数最多,而情感支持权重特别小或者特别大的用户很少,符合正态分布的基本特征.

7.3 对比算法

为了验证 MFP 算法的有效性,本文设计了多个对比算法和对比实验,综合验证了结构特征、行为特征和情感特征对意见领域挖掘效果的影响. 现将实验使用到的对比算法描述如下.

Rank 算法: 计算主题社区中用户节点的度,按度大小进行直接排序.

SRank 算法: 基于情感特征的排名算法(Sentiment-based Rank),将主题社区中用户获得的正向评价在获得的总评价数的比例定义为该用户的影响力.

HITS 算法: HITS 算法是网页链接分析中基础且重要的算法,它将每个网页节点的属性抽象为两种特征,一种特征是权威度 Authority,指与某个领域或者某个话题相关的高质量节点;另一种特征是中心度 Hub,指包含了很多指向高质量 Authority 节点链接的节点. HITS 算法为每个用户节点设定这两个属性,初始时都设定为 1,计算方法如式(7)所示:

$$\begin{aligned} Authority(v_i) &= \sum_{(V_j,V_i)\in E} Hub(v_j) \\ Hub(v_i) &= \sum_{(V_i,V_j)\in E} Authority(v_j) \end{aligned} \tag{7}$$

PageRank 算法: PageRank 算法是网页重要性排名的算法,也是本文 MFP 算法的参考算法. PageRank 算法节点重要性计算公式如下所示:

$$INF(u_i) = (1-d) + d^* \sum_{j\in Followers(i)}^{|Followers(i)|} \frac{1}{N_j} INF(u_j) \tag{8}$$

SHITS 算法: 在 HITS 算法原始结构的基础上,将用户节点的情感倾向作为中心度和权威度计算的权重. 按式(9)计算加权中心度和权威度:

$$\begin{aligned} Authority(v_i) &= \sum_{(V_j,V_i)\in E} W_{ji}^* Hub(v_j) \\ Hub(v_i) &= \sum_{(V_i,V_j)\in E} W_{ij}^* Authority(v_j) \end{aligned} \tag{9}$$

其中,W_{ij}为用户 v_i 对用户 v_j 的意见权重.

MFP 去除结构特征(Behavior + Senti),按照式(10)计算节点的影响力.

$$INF(u_i) = (1-d)B(u_i) + d^* \sum_{j\in Followers(i)}^{|Followers(i)|} \frac{W_{ji}}{|Friends(j)|} INF(u_j) \tag{10}$$

MFP 去除行为特征(Structure + Senti),根据式(11)计算节点的影响力.

$$INF(u_i) = (1-d)S(u_i) + d^* \sum_{j\in Followers(i)}^{|Followers(i)|} \frac{W_{ji}}{|Friends(j)|} INF(u_j) \tag{11}$$

MFP 去除结构特征和行为特征(Senti),根据式(12)计算节点的影响力:

$$INF(u_i) = (1-d) + d^* \sum_{j\in Followers(i)}^{|Followers(i)|} \frac{W_{ji}}{|Friends(j)|} INF(u_j) \tag{12}$$

7.4 评价指标

为了更全面的说明本文提出的 MFP 算法的有效性,实验使用了两种评价指标.

(1) **支持率**(Support Rate,SR):支持率从情感角度衡量了意见领袖在主题社区中受支持的程度,其计算公式如下:

$$SupportRate(v_i) = \frac{|v_j|}{N} \tag{13}$$

其中 N 表示社区中的总用户数，$|v_j|$ 表示参与话题讨论的用户中给予用户 v_i 正向评价的数量.

(2) 重合度(Top Overlap Ratio): 以考虑情感因素和不考虑情感因素所得到的意见领袖集合的重合度作为评价指标,其计算公式如下:

$$TopOverlapRatio = \frac{|TopResult_{senti}(a_1) \cap TopResult_{nosenti}(a_2)|}{K} \quad (14)$$

其中 K 为挖掘出的节点数目,a_i 表示所使用的算法,$TopResult_{senti}$ 表示基于情感的网络输出的 Top-K 节点集合,$TopResult_{nosenti}$ 表示基于无情感网络输出的 Top-K 节点集合.

7.5 实验结果与分析

实验一　MFP、SRank 和 SHITS 三种算法的对比实验

本文首先比较 MFP、SRank 和 SHITS 三种算法的实验结果,如图 5 所示. 由图可以看出,基于多维特征的 MFP 算法和 SHITS 算法明显好于直接排名的 SRank 算法. 当选取的 Top-K 较小时,MFP 算法可以获得很高的支持率,该算法只需选择较少的意见领袖节点便可获得更多用户的支持. 对于社交网络舆情监控或者广告营销行业,这种算法效果具有很大的现实意义.

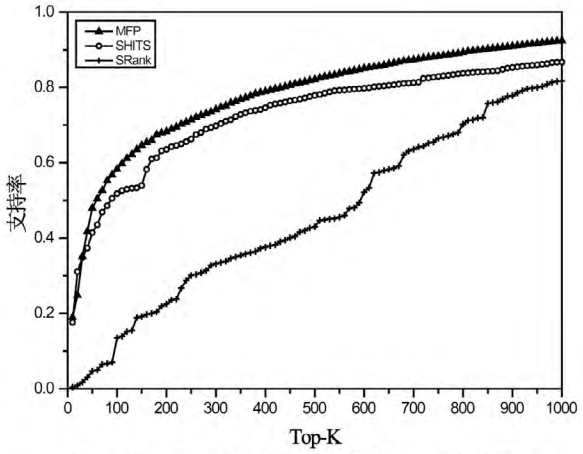

图5　MFP、SRank和SHITS三种算法实验结果比较

进一步分析可知,SRank 算法虽然考虑了情感特征,但只是对用户个体的影响力评估,而 MFP 算法综合考虑了周围好友的影响力,不仅考虑个体因素,而且考虑用户之间的潜在影响关系,这也印证了社交网络中用户的影响力相互作用的传播规律. 类似地,SHITS 算法使用权威度和中心度体现了用户之间的相互影响,相比 SRank 算法具有较大的优势.

实验二　MFP 算法去除行为因素、结构因素对比实验

为进一步研究结构特征和行为特征对节点影响力的影响,分别将 MFP 算法中的结构特征、行为特征去除,做出图 6 所示的对比实验. 其中,Structure + Senti 表示仅考虑结构特征和情感特征,Behavior + Senti 表示仅考虑行为特征和情感特征,Senti 表示同时去除结构特征和行为特征.

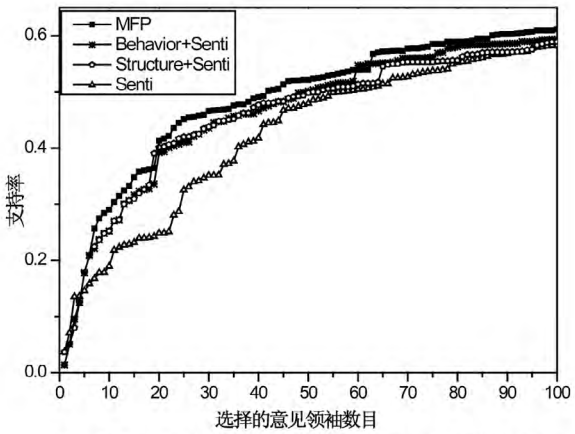

图6　MFP算法去除行为因素、结构因素对比实验结果

实验结果表明,当选取的意见领袖节点数 Top-K 较小时,加入了结构特征或者行为特征的算法实验效果比较接近,比不考虑两个特征的算法有明显优势,而同时考虑结构特征、行为特征和情感特征的 MFP 算法效果最好. 但是,随着意见领袖节点数 Top-K 的增大,结构特征和行为特征带来的优势逐渐消失. 可见在真实的社交网络中,结构特征与行为特征对于影响力较大的节点具有较高的区分度,但对影响力一般的用户节点,结构特征和行为特征并不是影响用户权威的主要因素.

实验三　MFP、PageRank、HITS 和 Rank 算法对比实验

此外,为了说明 MFP 算法在挖掘基于情感权重的意见领袖节点的有效性,本文首先通过实验将 MFP 算法和不考虑情感因素的 Rank、HITS 和 PageRank 算法做比较,以支持率作为实验效果的评价依据. 上述各算法均运行在 7.2 节描述的主题社区网络上. 实验结果如图 7.

从图中可以看出,虽然 MFP、PageRank、HITS 和 Rank 算法均可以用来挖掘传统意义上无情感的意见领袖节点,但是 MFP 算法相比其他算法能得到更高的支持率. 这说明按照传统意见领袖挖掘算法获得的节点难以体现社交网络“意见领袖”应具有的语义特征,虽然在社区中能影响到很多用户,但是没有真正起到领袖作用. 基于此,本文提出的 MFP 算法能够更精确地发现社交网络中的意见领袖.

实验四　MFP、PageRank 和 HITS 算法的重合度对比实验

本文最后结合重合度指标,统计得到 MFP 算法和

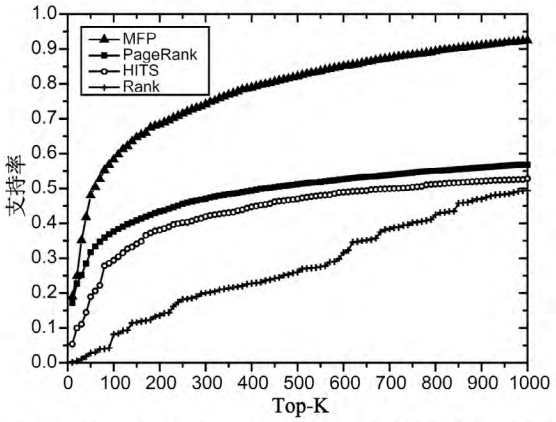

图7 MFP、PageRank、HITS和Rank算法对比实验结果

PageRank 算法、MFP 算法和 HITS 算法挖掘到的节点集合的重合度,实验结果如图 8 所示. 通过实验结果能够直观看出,随着选取意见领袖数目的增长,重合度逐渐增大并趋于平稳,社区中大部分高支持率的意见领袖都能被挖掘出来,未重合部分表明,MFP 算法能挖掘到传统算法容易忽视的节点,这些节点在传统意义上的影响力有限,但其言论得到本社区内其他用户的普遍认可,符合本文对于主题社区中意见领袖的定义.

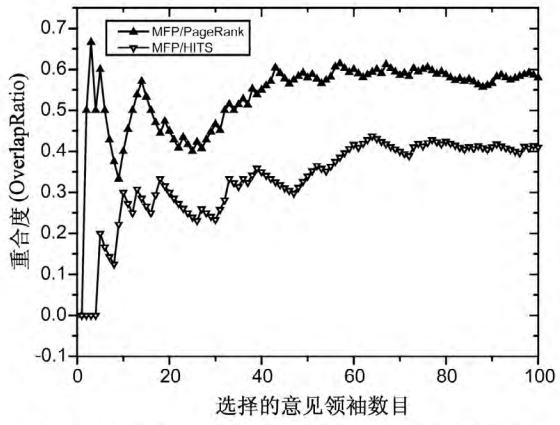

图8 MFP、PageRank和HITS算法的重合度对比实验

8 总结与展望

本文基于多维特征分析对社交网络中意见领袖节点挖掘问题进行研究. 从社交网络的拓扑结构与微博的语义内容出发,在考虑网络拓扑结构的基础上,兼顾动态特征以及用户对话题的感情倾向,更加准确地挖掘影响力节点. 多方面的对比实验表明,本文提出的 MFP 算法能有效挖掘出具有领域性的意见领袖节点,而且获得的意见领袖节点具有较高的支持率.

然而,本文仍然存在一些不足,如受新浪 API 的限制,无法获得用户的所有微博以及评论,因此针对内容的处理由于缺少规模较大的语料,准确率有待进一步加强. 此外,话题热度的上升带来的主题社区的规模不断壮大,随着社区规模的扩大,意见领袖识别算法的计算效率受到了严重的制约. 在以后的研究工作中,应发掘基于云计算平台的分布式意见领袖识别算法,提高其计算效率,从而提升意见领袖挖掘推广到企业应用的价值.

参考文献

[1] Ellison N B. Social network sites: Definition, history, and scholarship [J]. Journal of Computer-Mediated Communication, 2007, 13(1): 210 – 230.

[2] Kleinberg J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM (JACM), 1999, 46(5): 604 – 632.

[3] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [J]. ComputerNetworks and ISDN Systems, 1998, 30(1): 107 – 117.

[4] Kleinberg J M. Hubs, authorities, and communities [J]. ACM Computing Surveys (CSUR), 1999, 31(4es): 5.

[5] Zhai Z, Xu H, Jia P. Identifying opinion leaders in BBS [A]. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology [C]. IEEE, 2008. 398 – 401.

[6] Amit G, Francesco B, Laks V S L. Discovering leaders from community actions [A]. International Conference on Information and Knowledge Management (CIKM) [C]. California, USA, 2008. 499 – 508.

[7] Tsai M F, Tzeng C W, Lin Z L, et al. Discovering leaders from social network by action cascade [J]. Social Network Analysis and Mining, 2014, 4(1): 1 – 10.

[8] Xiaodan S, Yun C, Koji H, et al. Identifying opinion leaders in the Blogosphere [A]. Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management [C]. New York, USA, 2007. 971 – 974.

[9] Li Y, Ma S, Zhang Y, et al. An improved mix framework for opinion leader identification in online learning communities [J]. Knowledge-Based Systems, 2013, 43: 43 – 51.

[10] Zhou H, Zeng D, Zhang C. Finding leaders from opinion networks [A]. Intelligence and Security Informatics [C]. Dallas, USA, 2009. 266 – 268.

[11] Liu X, Wang Y, Li Y, et al. Identifying topic experts and topic communities in the blogspace [A]. Database Systems for Advanced Applications [M]. Berlin Heidelberg: Springer, 2011. 68 – 77.

[12] 夏霖. BBS 中组织拓扑结构研究和意见领袖识别 [D]. 武汉: 华中科技大学, 2011.
Xia Lin. Topology structure research and opinion leader identifying in BBS [D]. Wuhan: Huazhong University of

Science and Technology,2011. (in Chinese)
[13] Duan Jiangjiao,Jianping Zeng,Banghui Luo. Identification of opinion leaders based on user clustering and sentiment analysis [A]. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) [C]. IEEE,2014. vol 1.
[14] Song K,Wang D,Feng S,et al. Detecting opinion leader dynamically in Chinese news comments [A]. Web-Age Information Management [M]. Berlin Heidelberg: Springer,2012. 197 - 209.
[15] NTUSD. [EB/OL]. http://www.datatang.com/data/11837.
[16] HowNet. [EB/OL]. HowNet's Home Page. http://www.keenage.com.
[17] 同义词词林扩展版. [EB/OL]. http://www.ir-lab.org.

作者简介

曹玖新(通讯作者)　男,1967 年生于河南商丘,东南大学教授、博士生导师,主要研究领域为服务计算、网络安全、社会计算.
E-mail: jx.cao@seu.edu.cn

陈高君　男,1988 年生于河南漯河,东南大学硕士生,主要研究领域为社会计算.

吴江林　男,1988 年生于江苏南通市,东南大学硕士生,主要研究领域为社会计算.