ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 576－585
DOI:10. 16337/j. 1004-9037. 2022. 03. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多关系网络的话题意见领袖挖掘

段　震[1,2,3]，倪云鹏[1,2,3]，陈　洁[1,2,3]，张燕平[1,2,3]，赵　姝[1,2,3]

(1. 计算与信号处理教育部重点实验室，合肥 230601;2. 安徽大学计算机科学与技术学院，合肥 230601;3. 安徽省信息材料与智能传感重点实验室，合肥 230601)

摘　要：社交网络中的意见领袖在信息传播过程中起着重要的作用。传统的意见领袖挖掘仅基于网络结构,没有考虑特定话题或者事件下的作用,且目前基于话题的意见领袖挖掘仅基于单一的网络结构,并没有考虑到节点间的多种交互关系。本文提出一种基于多关系网络的话题意见领袖挖掘方法(Multi-relational networks，MRTRank),融合话题因素和节点间多种交互关系,通过一种属性网络表示学习算法,得到不同节点在多关系网络上的相似性,形成节点的转移概率矩阵,最终通过PageRank算法得到top-k个意见领袖。在真实Twitter数据集上的实验结果验证了本文提出的方法优于传统的意见领袖挖掘算法。

关键词：意见领袖;两级传播;社交网络;PageRank;属性网络表示学习

中图分类号：TP301.6　　**文献标志码**:A

Topic Opinion Leader Mining Based on Multi-relational Networks

DUAN Zhen[1,2,3], NI Yunpeng[1,2,3], CHEN Jie[1,2,3], ZHANG Yanping[1,2,3], ZHAO Shu[1,2,3]

(1. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Hefei 230601, China; 2. School of Computer Science and Technology, Anhui University, Hefei 230601, China; 3. Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei 230601, China)

Abstract: Opinion leaders in social networks play an important role in the process of information dissemination. The traditional mining of opinion leaders is based on network structures and doesnot consider the role of a specific topic or event, and the current mining of opinion leaders based on topic is only based on a single network structure, without taking into account the multiple interactive relationships between nodes. This paper proposes a topic opinion leader mining method based on multi-relational networks (MRTRank), which joins topic factors and a variety of interactive relationship between nodes. Through an attribute network representation learning algorithm, the similarity of different nodes in the multi-relationship network is obtained, and the transition probability matrix of nodes is formed. Finally, the top-k opinion leaders are obtained through the PageRank algorithm. Experimental results on real Twitter datasets verify that the proposed method is superior to traditional opinion leader mining algorithms.

Key words: opinion leader; two-step flow of communication; social networks; PageRank; attribute network representation learning

---

基金项目:国家自然科学基金(61876001);国防科技创新特区项目(2017-0001-863015-0009)。
收稿日期:2021-04-05;**修订日期**:2021-08-29

引　言

社交网络在人们的日常生活中起着重要的作用。一些在线社区或网站,如博客、维基和论坛等,人们可以很容易地分享他们的想法和观点,使得信息传递变得越来越频繁;社交网络的意见领袖指的是在特定领域内有影响力的人,有很多人关注他/她的评论或想法。所谓信息的"两级传播"是指信息首先从信息源头传播给意见领袖[1],然后由意见领袖将信息传播给普通大众。在社会网络中挖掘意见领袖具有巨大的商业和政治价值,通过确定最有影响力的人即可引导舆论。此外,发现最有影响力的评论也可以了解舆论的来源。目前,意见领袖的挖掘算法主要从节点的结构和属性两个角度出发提取用户的相似度信息,从而设计更加符合真实意义的PageRank[2]转移概率矩阵。很多基于PageRank改进的算法由于结合相关实际问题进行优化,相较于传统的PageRank,识别出的意见领袖更加精确。在社交网络中,用户间的交互行为已被证明是用户之间形成影响与被影响关系的一个重要因素,主要包括提及、点赞、转发和回复。这些交互行为说明了用户的紧密关系,因为它们明确地代表了用户对推文的反应,同时与推文的内容也具有强烈的相关性。例如,转发行为可以被认为是转发者向社交网络的用户发布其感兴趣的原始推文行为,从而潜在地增加原始推文的影响。在社交网络上,不同类型的交互行为都包含有用的信息,可以用来更好地衡量用户的影响力。

本文研究如何在社交网络中,通过用户的多种交互行为和话题偏好挖掘意见领袖。研究重点是如何使用用户交互行为和话题偏好来度量用户间的相似度,并据此设计转移概率矩阵,借助PageRank算法衡量用户的社交影响力。受如今网络中节点表示成功应用的启发,本文将不同交互网络的节点嵌入到向量空间中,并定义节点的紧密关系,它度量了节点在不同交互网络下的相似度。通过节点在多关系网络上的紧密关系和节点的话题偏好,设计节点的转移概率矩阵,通过PageRank算法为每个用户分配一个影响值,最终筛选出前top-k个用户作为意见领袖。在真实的Twitter网络数据集上的实验结果表明,由于引入了多关系网络和节点的话题偏好,本文提出的方法相较于传统的基于单一结构的算法,取得了更好的意见领袖挖掘效果。

1　相关工作

在意见领袖挖掘算法中,已有的方法主要分为3种:第1种通过使用社交网络的拓扑结构(主要是节点的关注关系网络)来捕获节点相对于整个网络的中心位置信息,以此衡量节点的影响力;第2种通过将节点的结构信息与节点自身的属性信息相结合,综合衡量节点的影响力;第3种通过节点间的多元关系衡量节点的影响力。使用拓扑结构信息进行意见领袖挖掘已有不少研究工作,主要从节点的低阶结构和高阶结构出发。使用低阶结构主要是利用节点的低阶邻居信息计算节点的影响力,例如常见中心性指标:度中心性、紧密度中心性等。高阶结构反映节点对在高阶邻居上的紧密程度。Zhang等[3]认为现有的研究主要集中在研究同阶邻居影响力,因此提出了一种新的结构影响力概念,研究了如何有效地从社交流中发现结构影响力模式。在微博数据集上的实验表明,与传统的影响模式挖掘算法相比,显著提高重要节点挖掘性能。Zhao等[4]提出传统的PageRank只使用基于边的关系,忽略了由模体捕获的高阶结构。Motif-based PageRank[5]在PageRank基础上使用多节点模体来捕获网络中节点之间的高阶关系,并引入两种融合方法挖掘基于节点间高阶关系的意见领袖。但是上述工作只是基于单一的网络结构,并没有考虑到节点在不同网络结构上的紧密程度。

除考虑节点的结构信息之外,经研究证明[6-8],考虑节点的属性信息能够提高意见领袖挖掘的准确率,如节点的话题偏好信息等,由此很多研究者提出了基于话题的意见领袖挖掘算法。这类算法认为

不同用户在不同话题下的影响力是不同的[6-12]，其中心思想是将用户发布的文本内容按话题进行具体划分，再在每个话题下进行意见领袖挖掘。对于不同的话题分配不同的权重，最后将不同话题下的用户影响力得分加权求和得到用户的总体影响力。TwitterRank[9]是其中最具代表性的意见领袖挖掘算法，其核心思想是用户的关注关系很大程度上取决于话题偏好的相似性。然而，TwitterRank忽略了用户之间的交互行为，只是基于单一的关注关系网络。除话题信息之外，社交网络中的用户有着各种属性信息，如粉丝数、发表推文数量、是否认证等，于是有研究者利用这些属性信息将社交网络用户群体进行分类[13-14]，如意见领袖用户、结构洞用户和普通用户等，然而这类算法只是将群体进行分类，并未计算用户具体的影响力值。由于关注关系网络只度量了节点的单一结构关系，不足以说明用户间真实的紧密关系，因此有研究者考虑在多关系网络上考虑节点的紧密关系。文献[15]算法考虑了微博用户自身因素与多种互动行为，如用户的粉丝数、用户活跃度、用户信用度、用户互动系数以及微博可见率等，利用改进的PageRank算法计算微博用户的影响力。文献[16]提出的EIRank使用节点嵌入方法将多种类型的交互网络集成到嵌入空间中，然后定义一种新的紧密度度量方法，以量化基于交互行为的用户间亲密度，但并未考虑这些交互行为与话题信息的关联性，即交互行为受到话题信息的影响。传统的意见领袖挖掘仅考虑单一的网络结构，忽略了话题偏好和用户的多种交互行为对于节点重要性的评估，即没有有效利用用户的多种行为特征，如转发、提及、点赞和评论等行为，而这些行为特征同样具有话题相关性[17]。因此本文结合节点在多关系网络上的信息和节点的话题偏好信息，利用PageRank算法评估节点的重要性。

2 相关定义

2.1 社交网络

定义1 社交网络$G=(V,E)$由1组节点V和1组边E组成，其中$E\subseteq V\times V$。如果$(m,n)\in E$，则m,n间存在1条边。如果(m,n)是有序的，则网络是有向的，否则是无向的。如果每条边都有不同的权重，则网络可以表示为$G=(V,E,\boldsymbol{W})$，其中$\boldsymbol{W}$为权重矩阵。

定义2 关注关系网络是1个有向社交网络，$G_{\text{following}}=(V,E_{\text{following}})$，表示用户及其邻居的关注关系。如果用户$m$被用户$n$关注，则存在1条边$(m,n)\in E_{\text{following}}$。

定义3 多关系网络是1个有向加权网络，$G_i=(V,E_i,\boldsymbol{W}_i)$，$i$表示用户间的交互类型，具体的交互类型包括转发、提及、回复、点赞。如果用户m、n存在交互行为，则存在1条边$(m,n)\in E_i$；$\boldsymbol{W}_i$表示用户间交互频率构成的矩阵。

2.2 加速属性网络嵌入

加速属性网络嵌入[18]（Accelerated attributed network embedding，AANE）是一种属性网络表示学习算法，可以把嵌入工作同时交给多个子任务独立、同时间完成，大大地节省了时间。对比一般的网络嵌入方法，如LINE、Node2Vec、DeepWalk等，AANE算法在很多数据集上表现出了优越性。AANE在分解结构信息矩阵的基础上，将属性信息也作为被分解的信息之一，使得矩阵分解能够同时受到结构信息和属性信息的约束，结构上距离相近结点的向量表示也应该接近，同时属性相似结点的向量表示也比较接近。并且，AANE将最终的优化问题转化为$2n$个子优化问题，采用分布式优化方式将原任务分解为多个子任务，子任务互相独立，因此大大降低了时间复杂度。本文中使用节点的多关系网络，在各关系网络中节点的属性信息各不相同。因此，提取节点在不同交互网络上的属性信息之后，采用并行的方式使用AANE提取多关系网络中节点的特征表示。

2.3 PageRank算法

PageRank是一种对网络中节点的重要性排序的算法,可以用来挖掘网络中的意见领袖。PageRank为每个节点分配1个初始排名(PR)值,其通常为$1/N$,其中N为节点的总数,通过节点的转移概率矩阵计算节点的影响力。经过若干次迭代,节点的PR值达到收敛状态。通常节点间的转移概率由节点的入度邻居数决定,对同一个节点的不同入度邻居,节点间的转移概率被均匀分配。然而对于不同用户之间,节点的紧密程度是不同的。因此,本文中节点间的转移概率由多关系网络上节点间的紧密关系和节点的话题偏好相似度共同决定。

3 本文方法

本文研究的问题是如何利用节点的话题偏好和多种交互网络信息来量化节点的社交影响力,即用户对邻居的影响不仅取决于用户间的话题相似度,而且取决于用户及其邻居有怎样的社交关系。本文的目标是设计一种统一的方法,利用节点的话题偏好(话题分布)和网络结构(交互网络信息)进行社交影响力排名,从而挖掘出top-k个意见领袖。

3.1 话题提取

对于用户话题偏好的提取,本文使用潜在狄利克雷分布[19](Latent Dirichlet allocation,LDA)模型。LDA模型是一种无监督的机器学习算法,用于从大型文档集合中识别潜在的话题信息。它使用了1个“单词袋”假设,将每个文档视为单词计数的向量。基于这个假设,每个文档在某些话题上被表示为文档-话题概率分布,而每个话题在多个单词上被表示为话题-词概率分布。对于每篇文档,话题的生成过程如下:(1)对每1篇文档,从文档-话题分布中抽取1个话题;(2)从上述被抽到的话题所对应的单词分布中抽取1个单词;(3)重复上述过程直至遍历文档中的每1个单词。

语料库中的每1篇文档与预先给定T个话题的1个多项分布相对应,将该多项分布记为θ。每个话题又与词汇表中的V个单词的1个多项分布相对应,将这个多项分布记为φ。

为了提取用户的话题分布,本文将用户发布的推文作为文档。由于推文字数的限制,LDA并不适合短文本话题建模[20],因此将每个用户发布的推文聚合到1个大文档中。对用户的文档进行话题提取即对应用户的话题偏好提取,每个文档本质上对应于1个用户。对所有用户的大文档进行训练得到的话题分布即对应用户的话题分布。

3.2 提取用户间紧密关系

本文提取用户紧密关系所使用的网络是节点间多种关系所形成的交互网络。直观上看,交互网络中边的权重是表示节点间紧密度的一种方法,但它不能度量未建立连接节点之间的关系,且用户的交互行为受各种因素的影响,每一种交互行为的影响因素各不相同[21]。因此将每一种交互行为的影响因素作为节点的属性,使用一种属性网络表示学习算法AANE,得到用户在多个交互网络上的向量表示,然后根据用户的向量表示得到交互网络中用户间的紧密关系。同EIRank算法[16],本文定义了4种向量空间:$\boldsymbol{S}=\{\text{Retweet},\text{Mention},\text{Reply},\text{Favorite}\}$,分别对应4种交互网络,如图1所示,其中权重表示用户之间交互的频率。因为节点间交互行为的影响因素各不相同,节点在各交互网络所具有的属性依具体交互网络而定。所使用到的属性包括节点的入度、节点的出度、发表的推文数量、节点粉丝数以及节点的话题偏好。 经过AANE得到各交互网络中节点的d维向量表示$\boldsymbol{Vec}_{\text{retweet}}$,$\boldsymbol{Vec}_{\text{mention}}$,$\boldsymbol{Vec}_{\text{reply}}$,$\boldsymbol{Vec}_{\text{favorite}}$,然后计算节点在每个向量空间的距离

$$\text{dist}(\boldsymbol{p},\boldsymbol{q})=\sqrt{\sum_{i=1}^{d}(q_i-p_i)^2} \tag{1}$$

式中$\boldsymbol{p}=(p_1,p_2,\cdots,p_d)$和$\boldsymbol{q}=(q_1,q_2,\cdots,q_d)$为节点$\boldsymbol{p}$和$\boldsymbol{q}$的$d$维向量。对于每个向量空间根据节点间的

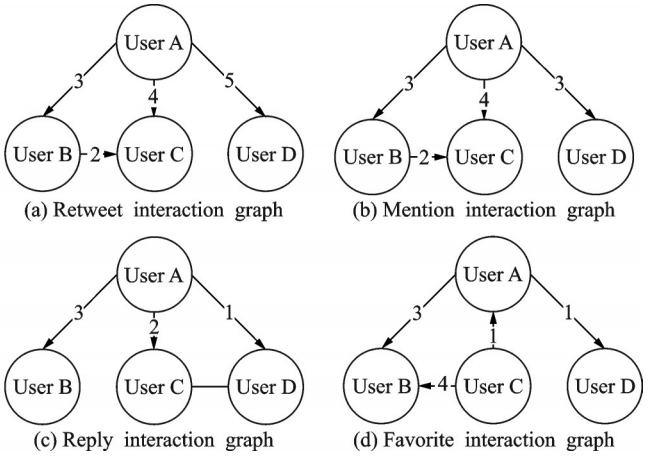


图1 4种交互图样例

Fig.1 Examples of four interaction graphs

欧式距离定义节点间的紧密关系,即

$$c_i(\boldsymbol{p},\boldsymbol{q})=\begin{cases}\dfrac{1}{\mathrm{dist}(\boldsymbol{p},\boldsymbol{q})} & \boldsymbol{p},\boldsymbol{q}\in \boldsymbol{Vec}_i, i\in S\\ 0 & \text{其他}\end{cases} \tag{2}$$

式中S为整个向量空间。

得到节点在各交互网络上的紧密关系之后,计算节点间的全局紧密关系。然而不同的交互网络对于节点的重要性贡献是不同的,因此分配权重给不同的交互网络,权重为某一种交互行为频率和总交互频率的比值,即

$$C(\boldsymbol{p},\boldsymbol{q})=\sum_{i=1}^{|\boldsymbol{Vec}|} w_i c_i(\boldsymbol{p},\boldsymbol{q}) \tag{3}$$

式中:$C(\boldsymbol{p},\boldsymbol{q})$表示节点间的全局紧密关系;$w_i$表示交互网络的权重,本文用到了4种交互关系,因此这里设置为1/4。

除了考虑节点在不同交互网络的相似性,话题意见领袖还须考虑节点间话题偏好的相似性,即

$$\mathrm{sim}_t(\boldsymbol{p},\boldsymbol{q})=1-\left|\boldsymbol{DT}_{pt}-\boldsymbol{DT}_{qt}\right| \tag{4}$$

式中$\mathrm{sim}_t(\boldsymbol{p},\boldsymbol{q})$表示节点$\boldsymbol{p}$、$\boldsymbol{q}$在话题$t$上的相似度,$\boldsymbol{DT}_{pt}$表示在文档-话题矩阵$\boldsymbol{DT}$中用户$\boldsymbol{p}$对话题$t$的话题偏好。

3.3 话题意见领袖挖掘

基于PageRank的改进算法主要思路是如何设计转移概率矩阵以描述用户间的紧密关系,以达到更精确的效果。从式(3,4)已得到节点间在各交互网络的紧密关系,然而节点在各话题下的影响力并不相同,因此设计节点在各话题下的转移概率矩阵为

$$\boldsymbol{TP}_t(\boldsymbol{p},\boldsymbol{q})=\frac{C(\boldsymbol{p},\boldsymbol{q})\cdot\mathrm{sim}_t(\boldsymbol{p},\boldsymbol{q})}{\sum\limits_{\boldsymbol{q}\ \text{follows}\ \boldsymbol{s}} C(\boldsymbol{p},\boldsymbol{s})\cdot\mathrm{sim}_t(\boldsymbol{p},\boldsymbol{s})} \tag{5}$$

式中:用户$\boldsymbol{p}$,$\boldsymbol{q}$为相邻关系;$\boldsymbol{s}$和$\boldsymbol{q}$为关注关系;$\boldsymbol{TP}_t$为在话题t下的转移概率矩阵,$\boldsymbol{TP}_t(\boldsymbol{p},\boldsymbol{q})$越高,说明用户$\boldsymbol{p}$受到用户$\boldsymbol{q}$影响的概率就越高;计算节点在每个话题下的影响力大小

$$P_{t,i}=\alpha\times TP_t\times P_{t,i-1}+\frac{1-\alpha}{N} \tag{6}$$

式中:P_t为在话题t下的影响力向量;α为阻尼系数;N为节点总数。

式(6)描述了节点在各话题下的影响力,通过聚合可以得到节点的综合影响力为

$$P=w_t\cdot P_t \tag{7}$$

式中:P为节点的综合影响力向量;w_t为分配给各话题影响力向量的权重。

为了挖掘top-k个话题意见领袖,本文给出一种基于PageRank的影响力计算算法,如算法1所示。

算法1　多关系网络话题意见领袖挖掘

输入:关注关系网络,用户交互网络集合G_i

输出:top-k个话题意见领袖

① for G_i in G:

② $\mathrm{Vec}_G=\mathrm{AANE}(G_i)$;

③ end for

④ 根据式(1~3)计算节点间的紧密关系;

⑤ 根据式(4~5)计算节点在每个话题下的转移概率矩阵;

⑥ 初始化每个用户的影响力值为$1/N$;/*N为节点个数*/

⑦ for t in all_topics:

⑧ 根据式(6)计算每个话题下的影响力向量;

⑨ end for

⑩ 根据式(7)计算节点的综合影响力向量,排序、筛选得到top-k个话题意见领袖。

4　实验及分析

4.1　数据集

实验所使用的数据集为公开数据集TIMME[22]。该数据集于2019年3月从Twitter收集,包含585位美国政客以及他们的关注者及后续关注者信息。数据集包括3个子集和1个总数据集。在整个推文数据集中提取相应用户的推文信息,去除停用词、低频词之后形成话题模型所需文本。经过预处理,得到的数据集共计5 435个节点,1 593 721条边,5种交互关系和9 548 310条推文。经处理后,数据集的信息如表1、2所示。其中表2具体展示了不同交互关系下的网络信息,包含了回复、转发、点赞、提及和关注关系,并将根据用户的所发文章提取了用户的话题偏好以及用户的属性信息,属性信息分别为具体交互网络下节点的出度、节点的入度、用户的粉丝数、所发文章数量以及话题偏好。

表1　数据集信息

Table 1　Information of datasets

网络	节点数	边数	推文数	节点间关系数
G_{TIMME}	5 435	1 593 721	9 548 310	5

表2　多关系网络信息

Table 2　Information of multi-relational networks

节点的关系网络	节点数	边数	提取话题数	使用属性数
G_{reply}	5 435	96 757	20	5
G_{retweet}	5 435	311 359	20	5
G_{favorite}	5 435	302 571	20	5
G_{mennton}	5 435	353 586	20	5
$G_{\mathrm{following}}$	5 435	529 448	20	5

4.2　实验结果

为了验证本文提出方法的有效性,本文选择以下常用的意见领袖挖掘算法作为对比:

(1) PageRank[2]:依据节点的关注关系度量

节点的重要性。

(2) Motif-based PageRank[5]:提取节点间的高阶结构关系-模体(motif),并据此设计转移概率矩阵。

(3) TwitterRank[9]:经典的话题意见领袖挖掘算法。其核心思想是,用户的关注行为很大程度上取决于用户间话题偏好的相似度。

(4) EIRank[16]:使用Node2Vec[23]得到节点在多种交互网络上的向量表示,根据节点向量表示的相似度设计转移概率矩阵。

意见领袖挖掘通常使用影响力传播模型来对算法进行评估。本文使用的影响力传播模型是线性阈值模型如图2所示。

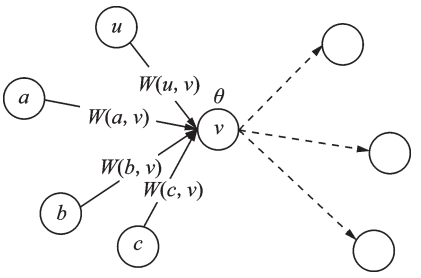


图2 线性阈值模型

Fig.2 Linear threshold model

在线性阈值模型中,每条有向边$(u,v)\in E$上都有1个权重$W(u,v)\in[0,1]$,直观上$W(u,v)$反映了节点u在节点v的所有入度邻居中影响力的重要性占比。每个节点v还有一个激活阈值$\theta\in[0,1]$,一旦确定在传播中就不再改变。在$s=0$时刻有且仅有种子集合中的节点被激活。当$s\geqslant 1$时,每个未激活节点v都需要依据它所有已激活的入度邻居来判断是否被激活,激活的判断标准为入度邻居对当前节点的影响力线性加权和是否已达到当前节点的激活阈值。若是,则节点v在时刻s被激活;否则,节点v仍然保持未激活状态。当某一时刻不再有新的节点被激活时,传播过程结束。本文选取各算法结果集中前top-k个节点作为种子节点集,并设置节点的激活阈值,通过种子节点集最终激活的节点数量作为评价指标。

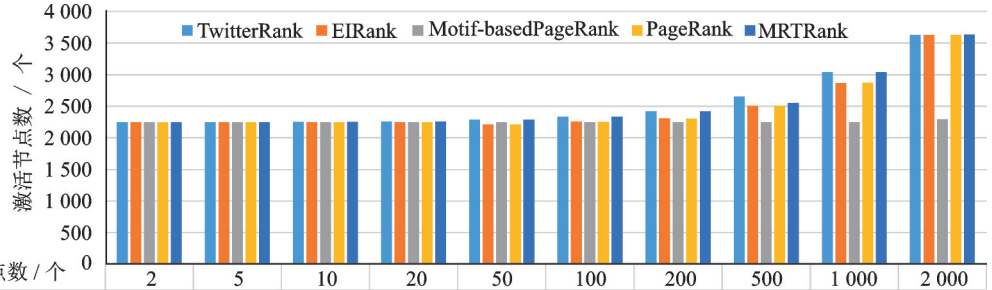
(1)用户在不同话题下的影响力各不相同,本节选取3个话题,每个话题下5个意见领袖作为种子节点,测试不同话题下意见领袖的影响力。TwitterRank和MRTRank的对比结果如表3所示。根据表3,从激活的节点数量看,两个算法性能接近;从具体激活的节点来看,激活的节点并不相同,这是因为一个用户对另一个用户产生影响具有话题相关性,即一个用户很大程度上受到一个与他话题偏好相似的用户的影响。在话题1中,意见领袖能够影响到的用户的话题偏好趋向于话题1,同理对于其他话题也是如此。

表3 部分算法结果对比

Table 3 Comparison of partial algorithm results

算法	话题编号	意见领袖节点	激活节点数量
TwitterRank	1	876、1 414、2 530、1 824、3 968	2 242
	2	6、5 247、327、5 034、1 637	2 243
	3	1 359、3 319、1 373、4 516、1 092	2 243
MRTRank	1	876、1 414、2 530、1 824、3 968	2 241
	2	5 370、3 474、221、3 843、3 059	2 243
	3	3 754、1 785、830、4 558	2 243

(2)各方法激活节点数量如图3所示。从图3可以看出,当选取种子节点数在20以内,各算法的最终激活的节点数量相差不多。这是因为各算法挖掘出的意见领袖大致相同,因此激活的节点数量相对一致。当选取种子节点数大于50时,各算法表现出现差异。部分算法中种子节点集中仍有意见领袖,所以激活节点数量保持着上升趋势,然而有些算法种子节点集中包含普通节点,所以激活的节点数量没有明显的变化趋势。Motif-based PageRank因为只考虑了单一的关注结构信息,没有考虑用户的多



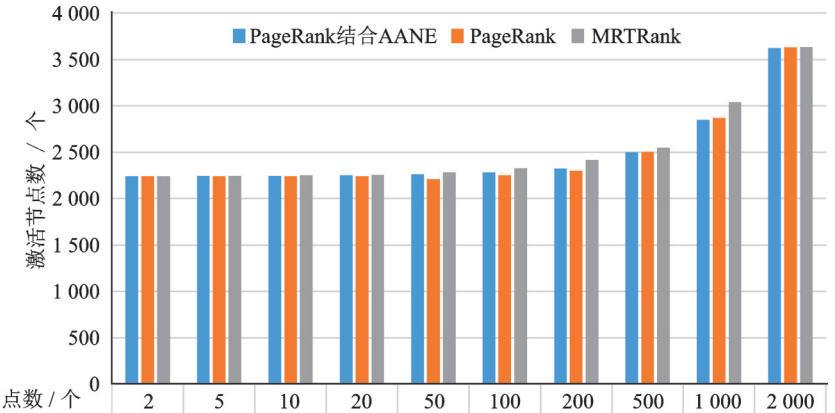
种子节点数 / 个	2	5	10	20	50	100	200	500	1 000	2 000
TwitterRank	2 244	2 247	2 251	2 257	2 285	2 329	2 417	2 649	3 036	3 631
EIRank	2 243	2 243	2 244	2 244	2 209	2 255	2 306	2 505	2 864	3 629
Motif-basedPageRank	2 243	2 243	2 243	2 243	2 243	2 243	2 243	2 243	2 243	2 292
PageRank	2 243	2 243	2 244	2 244	2 209	2 251	2 301	2 503	2 869	3 630
MRTRank	2 244	2 247	2 251	2 257	2 285	2 329	2 417	2 550	3 037	3 633

图 3　各方法激活节点数量

Fig.3　Number of activated nodes in each method

种交互关系以及用户的属性信息,随着种子数量的增多,相较于其他算法而言其激活的节点数量较少。EIRank考虑了多种交互信息,因此结果相对较好。MRTRank同时考虑了多种交互信息和话题偏好相似度,其结果与TwitterRank基本一致。当选取的种子节点为2 000时,除Motif-based PageRank算法之外,其他算法的实验结果基本一致,这是因为当种子节点达到一定数量时,各算法挖掘出的意见领袖节点几乎都在种子节点集中,因此最终激活的种子数量大致相同。从整体上看MRTRank和TwitterRank算法的表现力相差无几,两个算法都考虑了话题信息和用户的粉丝数量,研究这两个属性信息与用户的交互行为是否存在关联将是进一步的研究工作。

(3)对MRTRank进行消融实验,结果如图4所示。实验中主要考虑基于单一的关注关系网络、基于多种交互关系以及基于多种交互关系结合话题偏好相似度。图4结果表明,随着种子节点的增多,基于单一关注关系的PageRank算法效果较差,而PageRank结合AANE考虑了节点的多种交互关系,效果优于PageRank算法。MRTRank考虑节点多种关系并结合了节点间话题偏好相似度,效果最好。可以得出,除了用户的关注结构信息,用户的多种交互行为以及用户的话题偏好信息在意见领袖的挖掘



种子节点数 / 个	2	5	10	20	50	100	200	500	1 000	2 000
PageRank结合AANE	2 243	2 245	2 247	2 251	2 263	2 283	2 324	2 500	2 849	3 624
PageRank	2 243	2 243	2 244	2 244	2 209	2 251	2 301	2 503	2 869	3 630
MRTRank	2 244	2 247	2 251	2 257	2 285	2 329	2 417	2 550	3 037	3 633

图 4　消融实验结果

Fig.4　Ablation experiment results

中同样起到关键作用。

5 结束语

近年来,随着各种社交平台的兴起,人们之间的信息交流越来越频繁,使得意见领袖挖掘研究受到了广泛的关注。相对于传统的基于单一结构关系,本文所提出的方法将用户的话题偏好和多种交互关系相结合。首先收集用户的推文信息,并将用户推文合并成大文档,通过话题模型得到用户的话题偏好;其次构建4种不同的交互网络,并提取用户在具体网络上的属性信息,通过属性网络表示学习的方式,得到节点的向量表示,据此设计节点间的紧密关系;最终根据节点的话题偏好和各交互网络上的紧密关系设计转移概率,借助PageRank算法得到节点的影响力排名,选择top-k个影响力最大的节点作为意见领袖。在Twitter数据集上的实验结果证明本文提出的方法优于传统的意见领袖挖掘算法。

参考文献:

[1] ELIHU K. The two-step flow of communication: An up-to-date report on an hypothesis[J]. Public Opinion Quarterly, 1957, 21(1): 61-78.

[2] LANGVILLE A N, MEYER C D. Deeper inside pagerank[J]. Internet Mathematics, 2004, 1(3): 335-380.

[3] ZHANG Jing, TANG Jie, ZHONG Yuanyi, et al. StructInf: Mining structural influence from social streams[C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 73-79.

[4] ZHAO Huan, XU Xiaogang, SON Yangqiu, et al. Ranking users in social networks with higher-order structures[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2018: 232-239.

[5] ZHAO Huan, XU Xiaogang, SONG Yangqiu, et al. Ranking users in social networks with motif-based pagerank[J]. IEEE Transactions on Knowledge and Data Engineering, 2019(99): 1.

[6] 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(4): 490-493.
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(4): 490-493.

[7] HAVELIWALA T H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784-796.

[8] WENG Jianshu, LIM Ee-Peng, JIANG Jing, et al. Twitterrank: Finding topic-sensitive influential twitterers[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261-270.

[9] 曹林林, 郑明春. 微博话题符号网络下的意见领袖挖掘算法研究[J]. 计算机应用研究, 2017(12): 33-37.
CAO Linlin, ZHENG Mingchun. Algorithm of opinion leader mining based on signed network[J]. Application Research of Computers, 2017(12): 33-37.

[10] TANG Jie, WANG Chi, YANG Zi, et al. Social influence analysis in large-scale networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris:ACM, 2009: 807-816.

[11] QIAN Yang, LIU Yezheng, JIANG Yuanchun, et al. Detecting topic-level influencers in large-scale scientific networks[J]. World Wide Web, 2019, 23: 1-21.

[12] KIM D, LEE J G, LEE B S. Topical influence modeling via topic-level interests and interactions on social curation services [C]//Proceedings of 2016 IEEE 32nd International Conference on Data Engineering (ICDE). Helsinki: IEEE, 2016: 13-24.

[13] LIU Lu, TANG Jie, HAN Jiawei, et al. Mining topic-level influence in heterogeneous networks[C]//Proceedings of the 19th ACM Conference on Information and Knowledge Managemen. Toronto: ACM, 2010: 26-30.

[14] ZHAO Gouheng, JIA Peng, ZHOU Anmin, et al. InfGCN: Identifying influential nodes in complex networks with graph convolutional networks[J]. Neurocomputing, 2020, 414: 18-26.

[15] 王新胜, 马树章. 融合用户自身因素与互动行为的微博用户影响力计算方法[J]. 计算机科学, 2020, 47(1): 96-101.
WANG Xinsheng, MA Shuzhang. Method of weibo user influence calculation integrating users' own factors and interaction behavior[J]. Computer Science, 2020, 47(1): 96-101.

[16] BO H, MCCONVILLE R, HONG J, et al. Social network influence ranking via embedding network interactions for user recommendation[C]//Proceedings of Companion Proceedings of the Web Conference 2020. Taipei, China: ACM, 2020: 379-384.

[17] 韩忠明, 陈炎, 刘雯,等. 社会网络节点影响力分析研究[J]. 软件学报, 2017, 28(1): 84-104.
HAN Zhongming, CHEN Yan, LIU Wen, et al. Research on node influence analysis in social networks[J]. Journal of Software, 2017, 28(1): 84-104.

[18] HUANG Xiao, LI Jundong, HU Xia. Accelerated attributed network embedding[C]//Proceedings of the 2017 SIAM International Conference on Data Mining. Atlanta:Society for Industrial and Applied Mathematics, 2017: 633-641.

[19] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[20] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[C]//Proceedings of European Conference on Information Retrieval. Berlin, Heidelberg: Springer, 2011: 338-349.

[21] ELISA O, MANLIO D D, ALEX A. Characterizing interactions in online social networks during exceptional events[J]. Frontiers in Physics, 2015, 3(15): 59.

[22] XIAO Zhiping, SONG Weiping, XU Haoyan, et al. TIMME: Twitter ideology-detection via multi-task multi-relational embedding[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 2258-2268.

[23] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM, 2016: 855-864.

作者简介:

段震(1976-),男,讲师,研究方向:社交网络分析、社团划分、粒计算等,E-mail:ycduan@qq.com。

倪云鹏(1996-),**通信作者**,男,硕士研究生,研究方向:社交网络重要节点挖掘、影响力最大化等,E-mail:nyp_universe@163.com。

陈洁(1982-),女,副教授,研究方向:智能计算、机器学习等,E-mail:chenjie200398@163.com。

张燕平(1962-),女,教授,研究方向:智能计算、商空间、机器学习和智能信息处理等,E-mail:zhangyp2@gmail.com。

赵姝(1979-),女,教授,研究方向:网络表示学习、社交网络、粒计算等,E-mail:zhaoshuzs2002@hotmail.com。

(编辑:刘彦东)