基于情感倾向性分析的微博意见领袖识别模型

陈志雄[1] 王时绘[1] 高 榕[2]
(湖北大学计算机与信息工程学院 武汉 430062)[1] (武汉大学计算机学院 武汉 430072)[2]

摘 要 当前,微博意见领袖识别的研究方法纷繁多样,常见的方法有:对用户的个性化特征进行综合分析的方法和基于社交网络结构的分析方法。这些方法大多只考虑了用户的特征,未考虑用户之间的互动行为,或者未考虑微博文本的情感因素。为此,提出了一种基于微博情感分析的微博意见领袖识别方法。首先,在基于合成情感词典的词频统计结果的基础上,利用支持向量机对微博博文进行情感分析;然后,将变异系数法用于微博属性权重的计算,以体现微博的影响力;最后,利用改进的PageRank算法在微博用户转发关系网络中预测用户影响力的扩散过程,计算用户最终影响力的大小。在新浪微博数据集上通过实验评测该方法的性能,结果表明该方法能够有效提高识别性能。

关键词 微博,情感倾向性,变异系数法,意见领袖
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.05.028

Recognition Model of Microblog Opinion Leaders Based on Sentiment Orientation Analysis

CHEN Zhi-xiong[1] WANG Shi-hui[1] GAO Rong[2]
(School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China)[1]
(Computer School, Wuhan University, Wuhan 430072, China)[2]

Abstract The current methods of opinion leaders are numerous. Existing methods include the comprehensive analysis method for user's personalized features and the analysis method based on the structure of social networks. These methods only consider the characteristics of the users, but ignore the interaction between the users, or not take the microblog sentiment factors into consideration. Therefore, this paper proposed a microblog opinion leader identification method based on microblog emotional analysis. First of all, support vector machine is used to do emotional analysis on microblog context based on the synthetically emotional dictionary word frequency statistics results. And then the variation coefficient method is used in microblog attribute weight calculation to reflect the influence of microblog. Finally, the PageRank algorithm is employed to predict the diffusion process of users' influence in the microblog, and the final influence of users is calculated. On the Sina microblog dataset, the proposed model is tested and compared with the other methods based on user attribute analysis and PageRank using the forwarding rate and coverage index. The experimental results show the superiority of proposed method.

Keywords Microblog, Emotional tendency, Variation coefficient method, Opinion leader

1 引言

随着互联网技术的迅速发展,以博客技术为代表、围绕用户互动与个性体验的互联网应用技术,进一步推动了以开放、共享为特征的Web2.0时代向具有信息融合特征的Web3.0时代过渡。微博是一种具有代表性的技术,突出了互联网信息共享、网络信息传播的功能。它的广泛应用使得每位微博用户既能发布、创造各种信息,也能获取、处理各种信息,微博已经成为网络信息传播的主体。而在网络信息传播中,微博意见领袖作为网络信息的导向者,也凸显了微博作为在线社交媒体的属性特征。微博中的意见领袖活动在各类热点话题之中,他们观察问题细致入微,能够把握问题的核心,从而对问题的现象和本质做深入的剖析;更重要的是他们的剖析能够得到关注者的支持,并且影响关注人群的行为和看法,进而通过关注人群来传播这些观点。因此,意见领袖的影响力是不容忽视的。在实际应用方面,对意见领袖进行研究对于个人、企业和社会都有巨大的价值和意义。

目前,意见领袖的识别方法主要有两种:1)从用户的个人属性特征角度进行分析,构建意见领袖评估模型并对用户进行综合评分;2)从社交网络结构角度考虑用户之间的交互,对PageRank算法进行相应的改进,使其能够更有效地检测出意见领袖。虽然研究者从不同角度对意见领袖进行识别,并且

到稿日期:2017-02-06 返修日期:2017-05-06 本文受国家自然科学基金青年项目(41201404),国家自然科学基金项目(41401464,11101131),湖北省科技厅面上资助项目(2014CFB537)资助。
陈志雄(1983-),男,硕士,实验师,主要研究方向为人工智能、机器学习,E-mail:czxpro@163.com(通信作者);王时绘(1965-),男,硕士,教授,主要研究方向为人工智能、机器学习;高 榕(1981-),男,博士生,讲师,主要研究方向为人工智能、机器学习。

得出了很多研究成果，但是利用用户个性化特征或用户之间交互特征的识别方法进行建模的研究还存在不足，主要体现在如下两个方面：

1）通过直接获取量化的个性特征来评估用户的影响力大小，然后根据其影响力大小来判定该用户是否为意见领袖。虽然该方法比较直观且简单，但是其只考虑了用户特征的直接影响，而忽视了在网络信息传播过程中用户影响力的扩散性。例如，Cha 等[1]将时间因素和主题因素加入基于微博的转发次数、收听消息的人数和提及次数这 3 个指标中来度量博主的影响力，结果表明关注度较大即粉丝数量多的用户未必能获得较多的转发和评论数，但是影响力较大的用户在各类话题中占有重要位置。

2）基于社交网络链接结构的建模方法。该方法将 PageRank 等经典网页排序算法的思想应用到建模过程中，并加入主题因素、时间因素等指标来改进 PageRank 算法。虽然这类方法考虑了用户间的交互行为特征，但忽略了用户个人的情感倾向性的影响。例如，Weng 等[2]从主题相关的用户影响力方面进行分析并提出 Twitter Rank 算法，该算法对传统的 PageRank 算法进行了扩展，提取用户对所有关注人群的影响力，并将影响力之和作为该用户影响力大小的衡量标准。

针对上述研究中的不足，本文主要贡献如下：

1）提出了基于合成情感词典的微博情感特征词词频统计算法。该算法对微博进行中文分词，利用多个中文情感词典和由网络词汇合成的情感词典来统计各类情感词的词频。

2）将合成情感词典与支持向量机相结合，提出了改进的微博情感倾向性分析方法。该方法分析了微博博文内容的情感特征，采用合成情感词典来提取情感特征向量，并按照微博的情感倾向性利用支持向量机将微博分成了 3 类。

3）运用变异系数法进行微博属性权重的计算。该方法从数据中挖掘属性之间的联系，确定各个影响因素的客观权重，避免了人为因素的影响，对微博的初始影响力进行了量化。

4）改进了 PageRank 算法，将用户的所有微博初始影响力之和作为用户节点初始值，提出了基于转发概率的 PageRank 算法的微博影响力扩散模型。该模型在用户转发关系网络中预估影响力的扩散过程，计算微博传播后的用户影响力，得到最终影响力较大的用户（即意见领袖）。

5）在新浪微博数据集上对本文提出的模型进行实验，基于转发率、覆盖率将本文方法与基于用户属性分析的方法和基于 PageRank 的方法进行对比，实验结果表明基于情感倾向性分析的微博意见领袖识别方法能够对微博进行合理的情感分类，从而更准确地识别出转发率较高的微博意见领袖。

2　相关工作

2.1　情感倾向性分析的研究现状

微博情感倾向性分析就是将传统的情感分析方法应用于微博这种社交媒体中，对带有主观情感因素的微博文本内容进行综合分析、文本特征提取，进而研究情感分类的整个过程。目前，国内外的学者大多采用基于情感词典的方法和基于机器学习的方法进行情感倾向性分析。由于国外的研究起步较早，相关技术方法更为成熟，因此已有不少的研究成果。

Felipe 等[3]提出了基于元级特性情感分类的新方法，这种监督式的方法提高了基于 Twitter 文本的主观性和极性来检测情感分类的性能，而且对相应的研究结果进行主观性和极性的预测属于同一问题的不同方面，但它们需要使用不同的子空间功能来解决。Isidro 等[4]利用情感词典 HowNet 来识别文本中的情感词汇，采用基于领域本体和向量的计算方法来对意见进行分类。Gautam 等[5]将从文本中提取的形容词作为特征向量，选择特征向量列表，在 WordNet 上使用带有情感因素的支持向量机、最大熵以及朴素贝叶斯分类等 3 种基于机器学习的分类算法进行情感分析。同样地，Garg 等[6]在不同的公开数据集上运用朴素贝叶斯和最大熵两种分类方法来比较情感分析的准确度。Vanzo[7]将微博信息流中的情感分类问题定义为极性检测问题，采用 SVM 的方法进行情感极性分析。宋双永等[8]提出一种面向微博的热点事件情感分析方法。该方法首先自动挖掘用户对某热点事件的多个关注点，并针对不同关注点进行情感分析以及情感趋势监测，最终实现一个可视化的热点事件情感趋势分析原型系统。栗雨晴等[9]提出一种基于双语词典的多类情感分析方法，该方法通过构建双语多类情感词典对微博文本进行多分类语义倾向性分析。梁军等[10]主要探讨利用深度学习进行中文微博情感分析的可行性，采用递归神经网络来发现与任务相关的特征，并根据句子词语间前后的关联性引入情感极性转移模型以加强对文本关联性的捕获。张林等[11]研究了用户在智能设备上的评论，提取了这些短评论中的特征信息并筛选噪音，提高了情感分类的效果。刘志明等[12]采用多种机器学习算法，结合不同的特征选择方案，合理运用各种权重计算策略对微博情感分类方法进行深入挖掘，采用信息增益、SVM 以及 TF-IDF 相结合的方法作为特征项权重的计算方法，得到了最优的微博情感倾向性分类效果。

2.2　意见领袖的研究现状

近年来，意见领袖的研究引起了国内外学者的关注，他们得到了许多有价值的科研成果。Kwak 等[13]利用微博内容的转发次数和用户的粉丝数来度量用户的影响力从而判别意见领袖，结果表明该方法可以得到大量的评论和转发的微博信息，但是博主的粉丝数未必很多。Volpentesta 等[14]将商业社交网络定义为一个随时间变化的加权有向图模型，通过图中有向边随时间的变化来模拟社交网络节点之间的影响力变化，通过计算图中顶点的特征向量中心和时间价值中心来判别意见领袖。Momtaz 等[15]将意见领袖的所有属性分为结构、关系和个人特色，以识别意见领袖。曹玖新等[16]首先对微博真实文本数据进行话题识别从而得到主题社区，然后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，接着分别从结构、行为和情感 3 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。Zhang 等[17]通过分析系统公告中每篇文章的回复来提取社区，然后基于社区层次结构提出意见领袖社区挖掘方法。肖宇[18]综合考虑了聚类算法和分类算法的优势，提出一种基于话题内容分析的兴趣团体方法，在此基础上通过分析用户回帖的情感倾向来计算用户间链接的权重，最终提出了一种新的意见领袖发现算法。中科院的

马宁等[19]采用以网络拓扑结构与文本挖掘为基础的超网络理论进行意见领袖挖掘，提出了一种超网络模型（该模型分别从社会、心理、环境、观点这 4 层子网加以描述），并在该模型的基础上提出了一种新的超边排序算法（Super Edge Rank），该算法对用户间形成的超边进行排序，进而识别出意见领袖。樊兴华等[20]在影响力扩散概率模型 IDM 中引入有效词汇的概念，通过计算论坛帖子的影响因子，提出了具有开放性和包容性的网络意见领袖识别模型，提高了意见领袖识别的准确度。陈波等[21]提出了社交网络意见领袖的胜任力模型，该模型包括社交网络意见领袖应具有的信息生产、信息传播以及信息影响 3 大能力要素及各种显性和隐性行为指标。他们根据胜任力模型，将社交网络用户划分为普通大众、活跃分子、主题意见领袖和网络意见领袖 4 类，设计了意见领袖的层次筛选流程，从而最终实现意见领袖的识别。王晨旭等[22]基于社交网络中微博消息的传播与意见领袖的影响力密切相关，提出了一种基于消息传播的微博意见领袖影响力建模与测量分析方法；实验结果表明该模型能够较好地刻画意见领袖在消息传播过程中所起到的作用，并能够较为准确地对热门消息的传播趋势进行预测。

3 基于情感倾向性分析的意见领袖识别模型

本文主要是对特定话题下的微博内容和微博用户进行研究，从而建立相应的数学模型。首先提取微博内容的情感特征词汇，判别该条博文的情感倾向性；再将该博文划分到对应的博文类别中；然后将不同类别中的微博评论数量、转发次数、点赞次数、微博博主的个人影响力大小等因素相结合进行分析，利用变异系数法确定各种影响因素的权重，评估微博初始的传播影响力；最后计算转发网络中每个用户的初始传播影响力，预测用户转发网络中影响力的变化，得到影响力最大的用户，从而识别出对应类别的意见领袖。与其他以用户为中心、结合用户的个性化特征和行为特征并将其作为评价指标来进行建模的方法不同，本文把一段时间内特定话题下的微博作为研究对象，考虑了博文所属的情感类别，按照情感倾向对微博做分类研究，把微博的评论数量、转发次数、点赞次数、微博博主的个人属性等特征作为该微博的影响力评估指标，计算用户的微博初始影响力之和，预测微博用户的影响力变化，最终得到的影响力较大的用户即为对应话题下情感类别的意见领袖。

综上所述，基于情感倾向性分析的意见领袖识别研究的步骤如下：

1）对微博博文进行情感倾向性分析建模；

2）建立基于微博特征的影响力模型；

3）建立基于网络传播影响力扩散分析的微博意见领袖识别模型。

3.1 微博博文的情感倾向性分析

支持向量机分类方法（SVM）是一种基于结构风险最小化原理的新颖的机器学习算法，也是一种具有较好的泛化能力的预测工具，已经被广泛应用于文本分类以及人脸识别等领域。在文本分类领域，SVM 被证明是非常高效的，与传统的方法相比其鲁棒性更好[23-24]。

本文基于词典和机器学习的方法来对微博内容进行情感倾向性判别的建模。微博博文内容情感倾向性分析模型如图 1 所示。首先过滤掉微博中不必要的干扰信息；然后利用中文分词技术，根据合成后的情感词典查找文本中出现的情感特征词；再将情感词按类别组成文本的特征向量，并确定特征向量的权重；接着利用训练数据来训练 SVM 分类器；最后用该分类器进行情感分类。该方法不但降低了文本的特征空间向量的维度，而且排除了与情感倾向无关的特征，极大地提高了情感倾向性判别的效率和准确度。

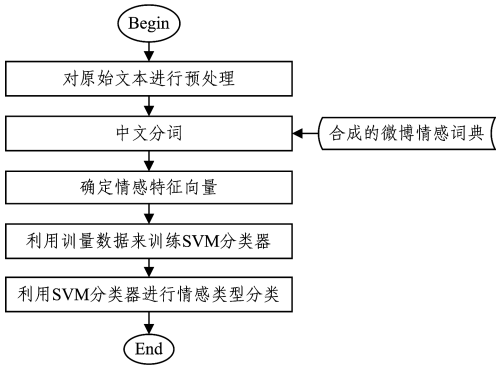

图 1 微博博文内容情感倾向性分析模型

Fig. 1 Sentiment orientation analysis model of microblog content

微博博文内容情感倾向性分析模型的相关步骤如下：

1）微博博文内容的预处理；

2）进行中文分词并合成微博情感词典；

3）确定情感特征向量；

4）利用 SVM 分类器进行情感分类。

其中，采用 SVM 支持向量机进行分类，其步骤如下：

1）确定特征向量，获取各个因素的特征值。

2）在训练样本上训练分类器。这里需要设计一个分类器，把微博按情感类别划分为带有正面情感的微博和带有负面情感的微博。

3）具有较大分类函数值的类别被划分为带有正面情感的微博；具有较小分类函数值的类别被划分为带有负面情感的微博；如果分类函数值介于最大值和最小值之间，那么可以认为其属于中性情感。

4）评估分类器的分类效果，并将该分类器运用到实际数据的分类中。

3.2 微博的属性特征计算

3.2.1 微博的属性特征分析

本文以微博为研究对象，通过评估微博的影响力大小来识别对应的意见领袖，首先需要确定与微博影响力有关的属性特征。对微博的传播和扩散起到决定性作用的因素有 3 个：社交网络自身结构的复杂性、微博信息的多样性和微博博主的个性特征。

本文将影响微博信息传播的特征（微博的评论数量、转发次数、点赞次数、微博博主的个人影响力大小）提取出来并进

行评估。微博的评论数量表示其他用户对某条微博的评论条数，该微博特征能在一定程度上体现其他用户受微博信息的影响程度，体现了微博信息对用户行为的影响。转发次数是指对博主发布的状态进行直接转发的人数。点赞次数则表示对博主发布的信息或者观点表示认同和支持的人数。考虑到用户的个人属性较多以及量化的难易程度，本文只将粉丝数、关注数、历史微博数这 3 个指标纳入决定微博博主的个人影响力因素的范围。若微博博主的粉丝数量越多，则其博文的用户评论和转发就越多。若关注的好友数量越多，则其获取信息的渠道越广。历史微博数量能在一定程度上体现用户的活跃度，用户活动越频繁，越可能成为意见领袖。博文的情感倾向性分析算法如算法 1 所示。

算法 1　博文的情感倾向性分析算法

输入：m 条微博博文的情感特征向量 WeiboVector

输出：m 条博文对应的情感类别

1. 将博文的特征向量 WeiboVector 转换为 libsvm 需要的数据格式：Catalog 1：t_1 2：t_2 3：t_3 4：t_4 5：t_5；
2. 对样本数据 WeiboVector 进行训练 train(WeiboVector)，得到模型 modelFile；
3. 利用训练得出的模型 modelFile 对测试数据进行分类预测，从而得到：resultFile：predict(WeiboVector，modelFile，resultFile)；
4. 计算预测结果的准确率 accuracy

该算法的时间复杂度、空间复杂度均与 $m * n$ 阶的情感特征矩阵有关，其时间复杂度和空间复杂度均为 $O(m * n)$。该分类器具有较高的性能，可以用作情感倾向性分析的分类器。

3.2.2　微博属性特征值计算模型

在进行多因素分析时，需要确定各个因素的权重。本文采用的是基于客观数据的计算方法——变异系数法。基于客观赋权的变异系数能够减少人为因素的干扰，更加客观地分配各个权重，进而对属性特征值进行量化计算[25]。微博的节点属性特征值计算模型如图 2 所示。

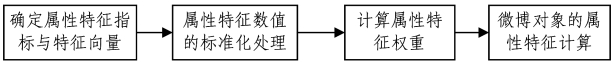

图 2　微博节点的属性特征值计算模型

Fig. 2　Attribute feature computational model of microblog spreader

该模型主要包括以下 4 个步骤：

1）确定属性特征指标与特征向量

待评估的对象就是获取的微博对象的集合 $U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$，每条微博的属性特征指标可定义为属性特征向量 $e=\{e_1,e_2,e_3,e_4,e_5,e_6\}$，$e_1$ 表示对该微博的评论数量，e_2 表示转发次数，e_3 表示点赞次数，e_4，e_5，e_6 分别表示博主的粉丝数、关注数和历史微博数。

2）属性特征值的标准化处理

获取到的数据具有不同的参考标准，如微博的评论数量、转发次数、点赞次数属于对用户的行为特征的量化，博主的粉丝数、关注数、历史微博数属于博主的个性化特征。对原始的数据进行归一化处理有利于排除上述参考标准的干扰。微博节点的属性特征值计算模型使用的是线性标准化的方法——min-max 标准化，该方法最终把数值变换到[0，1]区间。

原始的数据矩阵为 $D=\{d_{i1},d_{i2},d_{i3},d_{i4},d_{i5},d_{i6}\}$，$i\in N$，经过标准化处理后为 $S=\{s_{i1},s_{i2},s_{i3},s_{i4},s_{i5},s_{i6}\}$，$i\in N$，定义 $d_j^{\max}$ 为原始矩阵中第 j 列的最大值，定义 $d_j^{\min}$ 为原始矩阵中第 j 列的最小值。通过线性标准化的方法得：

$$S_{ij}=\frac{d_{ij}-d_j^{\min}}{d_j^{\max}-d_j^{\min}} \tag{1}$$

3）计算属性特征的权重

定义 V_i，σ_i，$\overline{X_i}$ 分别为第 i 个因素的变异系数、标准差、平均数。基于标准化处理后的数据，计算各个影响因素的变异系数，计算公式如下：

$$V_i=\frac{\sigma_i}{\overline{X_i}},i=1,2,\cdots,n \tag{2}$$

根据计算结果可以得到影响因素集合对应的权重。

4）进行微博对象的属性特征计算

对第 i 条微博的属性特征进行量化，计算公式如下：

$$I_i=\sum_{i=1}^{n}w_j * s_{ij},i\in N,j\in N \tag{3}$$

该计算结果是对单条微博信息的初始影响力的度量。

利用变异系数法计算微博初始影响力的算法如算法 2 所示。

算法 2　利用变异系数法计算微博初始影响力的算法

输入：某类微博中的 n 条微博的属性特征向量 propertyVcetor

输出：n 条微博的初始影响力大小

1. 构建微博的属性特征矩阵 E=$\{E_1,E_2,\cdots,E_6\}$；
2. 对微博的属性矩阵进行归一化处理得到矩阵 S←process(E)；
3. 计算矩阵 S 各列的标准差 $stdev_j$ 以及平均值 avg_j；
4. 计算矩阵 S 各列的变异系数 $V_j\leftarrow\frac{stdev_j}{avg_j}$；
5. 计算各个属性对应的权重 $w_j\leftarrow\frac{v_j}{\sum_{j=1}^{n}v_j}$；
6. 计算每条微博的初始影响力 $I_j\leftarrow\sum_{j=1}^{n}w_j * S_{ij}$。

算法 2 的时间复杂度与微博的数量相关，因此其时间复杂度为 $O(n)$。空间主要用于存储中间计算结果，因此空间复杂度也为 $O(n)$。

3.3　微博意见领袖识别模型

微博信息主要沿着用户转发关系网络中的路径不断扩散。若要识别出影响力较大的用户，则需要计算研究范围内的所有用户在社交网络中的影响力大小，实际上就是评估该用户在网络中的排名，因此采用 PageRank 算法在用户转发网络中预测用户影响力的变化，当网络稳定后，该排序算法可以根据微博用户的影响力大小对用户进行排序，影响力最大的用户即为意见领袖。

在本文提出的模型中，每个微博信息的初始值为 3.2 节中计算出来的微博属性特征值。由于每条微博的属性特征值都不同，而同一个用户可能发布了多条微博，该用户的初始影响力为所有微博的初始值之和，因此微博转发网络中每个用户的影响力的初始值也不同。此外，考虑到微博转发过程中用户之间的互粉关系和历史转发记录，将转发概率引入到

PageRank 算法中以得到更为合理的计算结果。但经典的 PageRank 算法设置了一个相同的初始 *PR* 值，如果直接用其来建模，则得到的结果不准确，而且必须充分考虑该算法的收敛条件。常见的收敛标准有 3 种：1)每个网页的 *PR* 值与上一次计算的 *PR* 值相等；2)当所有网页的 *PR* 值与上一次计算的 *PR* 值的差值平均小于某个差值标准时，认为算法收敛；3)一定比例的网页的 *PR* 值和上一次计算的 *PR* 值相等。本文采用第二种收敛标准作为算法的结束条件。当算法收敛时，各个用户节点的影响力值趋于稳定，此时影响力较大的用户即为意见领袖。改进的 PageRank 算法如算法 3 所示。

算法 3 改进的 PageRank 算法

输入：m 条微博的初始影响力大小

输出：n 个用户的最终影响力大小

```
1. for(i=0;i<m;i++)
2.    user=findUser(mid);//通过微博的 mid 找到对应的用户
3.    if(!userMap.contains(user.getUid()))//如果结果集中不包含
      该用户
4. userMap=add(user.getUid(),user.getPow());
5.    else //若存在该用户则更新初始影响力
6.       userMap=update(uid,pow);
7.    endif
8. endfor
9. M=buildForwardMatrix(userMap);//构建用户转发关系矩阵
10. F=calculateForwardRate(userMap);//计算用户之间的转发率
11. S=calculates(M,F);//计算初始的用户关系矩阵
12. g=addMatrix(S,userMap.Size());//计算 G 矩阵
13. List<Double> q;
14. while(true){
15.    q=vectorMulMatrix(g,q1);//矩阵 g 和 q 相乘
16.    flag=checkDistance(q,q1);//检查矩阵 g 和 q 所有元素的差值
17.       if(flag == true) break;
18.       endif
19.       q1=q;
20. }
21. return q;
```

PageRank 算法不断地进行迭代计算，本文将矩阵 G 和 R 相乘，不断地重复该过程，直至矩阵 R 中的所有元素与上一次相乘的结果的差值都小于设定的阈值时算法收敛。算法收敛时，矩阵 R 中的 n 个值对应 n 个微博用户节点的最终影响力值。最后对用户节点的最终影响力大小进行比较，影响力较大的用户即为意见领袖。

该迭代计算的过程可以看作是用户影响力沿着用户转发关系网不断扩散的过程。本文将用户的初始影响力作为节点的初始值，利用转发概率来分配用户的影响力，改进了传统 PageRank 算法。使用该改进算法来模拟用户的影响力扩散过程，其时间消耗与用户的数量 n 有关，空间主要用于存储中间计算结果，因此该算法的时间复杂度和空间复杂度都为 $O(n^2)$。

4 实验

4.1 数据集

本文实验基于新浪微博数据集。新浪微博数据集是利用新浪微博为开发者提供的 Sina API 从新浪微博上爬取得到的。如表 1 所列，该数据集包括 63641 条新浪微博用户信息，1391718 条用户好友关系，27759 条微博转发关系以及 84168 条在 2014 年 5 月 3 日至 2014 年 5 月 11 日间关于 12 个主题的微博信息。

表 1 数据集概要

Table 1 Information of dataset

数据名称	数据
新浪微博用户信息数	63641
用户好友关系条数	1391718
微博转发关系条数	27759
主题微博信息数	12
主题微博信息抓取时间	2014 年 5 月 3 日—2014 年 5 月 11 日
主题微博信息转发条数	84168

4.2 实验环境

实验环境参数如表 2 所列。

表 2 实验环境参数

Table 2 Parameters of experimental environment

实验设备	实验环境
CPU	Intel Core i5-4258 2.40 GHz 双核
内存	12GB
操作系统	Windows 8.1(64 位)
软件开发环境	MyEclipse 8.5+JDK1.7
数据库	MySql 5.3

4.3 评估标准

由于目前还没有一个公认的意见领袖评判标准，单独地根据粉丝数、转发数来判定意见领袖是不准确的，因此本文根据意见领袖在社交网络中产生的影响力范围，采用微博转发率、用户覆盖率两个指标的数值大小来评估意见领袖判定结果的准确性，并且对比基于用户属性分析的方法和基于在用户关注网络中应用 PageRank 算法的识别方法来验证模型的有效性和合理性。

微博转发率(Forwarding rate)是指某用户在一段时间内发布的所有微博转发数量占所有用户发布的微博转发数量的比例。令某用户的所有微博的直接转发数量为 N_d，间接转发数量为 N_j，所有用户的微博转发总数为 N_t，则转发率的计算公式下：

$$Forward=\frac{N_d+N_j}{N_t} \tag{4}$$

用户覆盖率(Coverage rate)是指在用户交互网络中受意见领袖影响的用户数占该网络中所有用户数的比例。记前 k 个意见领袖影响的用户数量为 N_k，用户总数为 N_s，则用户覆盖率的计算公式如下：

$$Coverage(k)=\frac{N_k}{N_s} \tag{5}$$

4.4 实验分析

本文结合知网、大连理工大学和台湾大学等多个研究机

构对情感词典的研究成果，以及“网词网”[1] 对最新网络词汇的整理结果，并根据微博博文内容的特有情感特征合成微博的情感词典。该情感词典的组成如表 3 所列。

表 3　微博情感词典的组成

Table 3　Construction of microblog sentiment lexicons

词汇来源	正面词语或正面表情个数	负面词语或负面表情个数
知网	4566	4370
大连理工	11229	10783
台湾大学	2810	8276
网词网	26	51

根据表 3 对微博情感词典的组成进行分析，统计得到合成后的情感词典中正面情感词为 18631 个，负面情感词为 23480 个，正面表情符号为 25 个，负面表情符号为 28 个，常用否定词为 26 个。

本文情感特征向量 $T=T(t_1,t_2,t_3,t_4,t_5)$。其中，t_1，t_2，t_3，t_4，t_5 分别表示正面词语个数、负面词语个数、正面表情个数、负面表情个数和否定词个数。

从 2014 年 5 月 3 日至 2014 年 5 年 11 日的微博中选择了 1514 条待评估的微博，过滤掉评论数量较小的微博，最后对 224 条微博进行研究。对这些微博的情感倾向性进行分析，得到含有 95 条持否定态度的微博的集合 C_1，77 条持肯定态度的微博的集合 C_2，52 条持中立态度的微博的集合 C_3。

同时，基于以上情感词典，对 224 条微博的情感词词频进行统计，结果为：正面情感词 1462 次，负面情感词 1326 次，正面表情符号 206 次，负面表情符号 242 次，否定词 95 次。

定义微博的属性特征矩阵 $E(E_1,E_2,\cdots,E_6)$，E_1 表示对微博的评论数量，E_2 表示微博的转发次数，E_3 表示微博的点赞次数，E_4，E_5，E_6 分别表示博主的粉丝数量、关注数、历史微博数量。基于式(2)，利用变异系数法求解特征属性的权重，权重统计结果如表 4 所列。

表 4　基于微博集合的变异系数得到的权重统计结果

Table 4　Weight statistical results of variation coefficient based on microblog set

	E_1	E_2	E_3	E_4	E_5	E_6
C_1 变异系数	2.1247	1.6361	1.9493	5.9605	0.6104	1.4967
C_1 权重	0.1542	0.1187	0.1415	0.4326	0.0443	0.1086
C_2 变异系数	2.6256	3.4788	1.5950	2.2447	0.4895	0.5569
C_2 权重	0.2389	0.3165	0.1451	0.2042	0.0445	0.0507
C_3 变异系数	1.9407	1.6381	2.1070	3.1415	0.6902	1.0567
C_3 权重	0.1835	0.1549	0.1993	0.2971	0.0653	0.0999

同时，为了验证本文所提出的基于情感倾向性分析的意见领袖识别方法的有效性，在相同的新浪微博数据集上分别采用基于量化的用户属性值的方法以及基于 PageRank 算法的识别方法进行对比实验，采样数据如表 5 所列。基于量化的用户属性值的方法主要将与用户影响力相关的属性进行加权求和，考虑了粉丝数、评论数量、转发次数、点赞次数，使用的是文献[26]中的方法。而基于 PageRank 算法的意见领袖识别方法使用的是文献[13]中的方法，即在用户间的关注关系网络中计算用户的影响力。

[1] http://wangci.net

表 5　3 种方法所得的意见领袖微博转发率采样数据

Table 5　Opinion leader microblog forwarding rate of sample data among three methods

序号	用户编号	用户 uid	微博转发率/%		
			用户属性值方法	PageRank 方法	本文方法
1	u_1	1618051664	3.75	3.75	3.75
2	u_2	2656274875	0.81	—	0.81
3	u_3	2286908003	0.15	—	—
4	u_4	1893801487	0.63	—	—
5	u_5	1638782947	0.98	—	0.98
6	u_6	1699432410	0.51	—	—
7	u_7	1649173367	0.14	—	—
8	u_8	1642088277	0.90	—	—
9	u_9	1644489953	0.30	—	—
10	u_{10}	2032139271	0.09	—	—
11	u_{11}	1266058190	—	1.90	1.90
12	u_{12}	2165372481	—	2.65	2.65
13	u_{13}	1885454921	—	3.00	3.00
14	u_{14}	3512454351	—	4.03	4.03
15	u_{15}	1743951792	—	2.18	2.18
16	u_{16}	2111717203	—	7.20	7.20
17	u_{17}	3276353050	—	3.46	3.46
18	u_{18}	1093974672	—	3.52	3.52
19	u_{19}	3209903582	—	3.19	3.19

4.4.1　基于微博转发率的对比分析

将本文方法与基于用户属性值分析的识别方法(方法 1)进行对比分析。本文提出的方法得到了 3 类意见领袖，共有 12 个博主，其对应的用户 uid 分别为：“1618051664”“2656274875”“1266058190”“2165372481”“1638782947”“3512454351”“1743951792”“2111717203”“3276353050”“1093974672”“3209903582”“1885454921”。本文方法提取前 5 位意见领袖，其对应的用户 uid 分别为：“1618051664”“2656274875”“1266058190”“2165372481”“1638782947”。而在方法 1 得到的前 10 位意见领袖中，与本文方法得到的意见领袖相同的用户只有 3 位，其对应的用户 uid 分别为“1618051664”“2656274875”“1638782947”。根据用户转发率(式(4))，可以得到图 3 所示的计算结果。

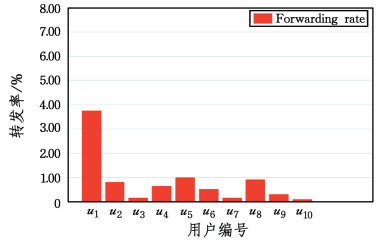

图 3　基于用户属性值方法所得的意见领袖微博转发率

Fig. 3　Opinion leader microblog forwarding rate based on user attribute value method

根据图 4 和图 5，将基于 PageRank 算法的意见领袖识别方法(方法 2)与本文方法进行对比分析，所得的意见领袖的转发率分别如图 4、图 5 所示。本文方法得到的意见领袖包括通过方法 2 识别出的博主，而且还对博主观点的情感倾向性进行了分类，有利于发现不同情感倾向的舆论引导者。例如，uid 为“1638782947”的用户发布了转基因作物广泛种植和

食品流通的博文，表明了他对“转基因”的肯定；uid 为“1649159940”的用户发布了对转基因的争议博文，没有特定的情感倾向性，只是客观地阐述问题，但也有一定的影响力。由此可见，与方法 2 相比，本文提出的方法能对微博用户的情感倾向进行分类，能挖掘出更多特定情感倾向的意见领袖，并且还可以按照情感倾向对意见领袖进行类别划分。

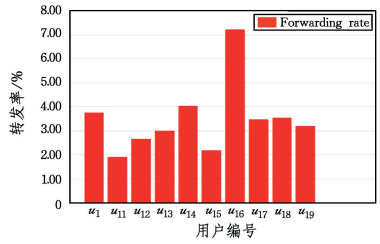

图 4 基于 PageRank 方法所得的意见领袖微博转发率

Fig. 4 Opinion leader microblog forwarding rate based on PageRank method

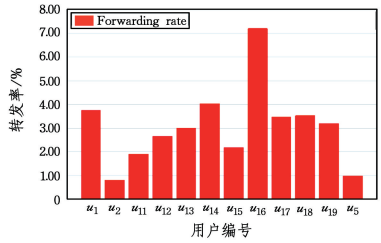

图 5 本文方法所得的意见领袖微博转发率

Fig. 5 Opinion leader microblog forwarding rate based on our method

4.4.2 基于用户覆盖率的对比分析

意见领袖影响的用户范围越广，则说明识别意见领袖的方法越有效。4.3 节使用用户的微博转发率来评估意见领袖的识别结果，本节将使用前 k 个意见领袖影响的用户比例来评估 3 种识别方法。根据式(5)得到 3 种方法对应的前 k 个意见领袖的用户覆盖率，如图 6 所示。

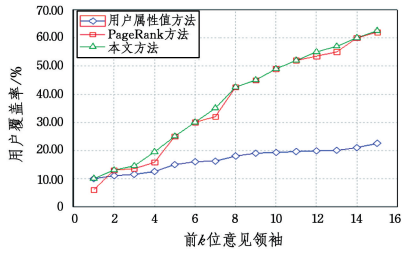

图 6 基于用户覆盖率的 3 种方法的对比

Fig. 6 Comparison among three methods based on user coverage

从图 6 中可以看出，方法 2 与本文方法比方法 1 的用户覆盖率更高。而在前 10 位意见领袖的影响下，方法 2 与本文方法覆盖的用户范围非常接近。在前 10～15 位意见领袖的影响下，本文方法覆盖的用户范围更广。这充分说明了本文提出的识别方法不但能识别出相应的意见领袖，而且这些意见领袖影响的用户范围更广。

综上所述，与用户属性值方法和 PageRank 方法相比，本文提出的基于情感倾向性分析的意见领袖识别模型能够按照情感倾向性对意见领袖进行划分，而且意见领袖影响的用户范围更加广泛，在一定程度上提升了识别准确度。因此，所提模型对意见领袖的识别是有效的。

结束语 本文将微博博文的情感倾向性因素加入意见领袖识别模型中，以新浪微博数据为研究对象，提出了基于情感倾向性分析的意见领袖识别方法。本文主要完成了以下研究工作：

1)将情感词典与支持向量机相结合，改进了微博情感倾向性分析的方法。基于合成情感词典的词频统计结果，利用支持向量机对微博博文进行情感类别分析，分析该博文对事件或者话题的情感倾向，从而将博文分为持肯定态度的博文、持否定态度的博文以及持中立态度的博文。

2)将变异系数法运用于微博属性权重计算。该方法利用微博的评论条数、点赞次数、转发次数、博主的粉丝数量、博主的关注数、博主的历史微博个数等 6 个因素构建微博影响力度量指标矩阵，并采用变异系数法计算各个因素的客观权重，避免了主观因素对权重的影响，同时计算出了微博的初始影响力。

3)提出了基于改进的 PageRank 算法的微博影响力扩散模型。该模型将用户的多条微博的影响力之和作为用户节点的初始迭代值，利用 PageRank 算法在微博用户转发关系网络中预测用户影响力的扩散过程，计算用户最终的影响力大小。通过与用户属性值方法、PageRank 方法进行对比分析，验证了本文所提出的基于情感倾向性分析的微博意见领袖识别方法的有效性和优越性。

未来的工作将考虑时间因素，包括用户发布微博的时间规律、微博的评论和转发行为随时间的变化趋势。同时，我们将研究如何自动识别网络流行词汇并将其加入到情感词典中，以提高情感分析算法的准确度，从而进一步提高识别的准确率。

参 考 文 献

[1] CHA M, HADDADI H, BENEVENUTO F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy[C]// Proceeding of 10th International AAAI Conference on Weblogs and Social media(ICWSM). 2015: 10-17.

[2] WENG J, LIM E P, JIANG J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining(WSDM). 2010: 261-270.

[3] BRAVO-MARQUEZ F, MENDOZA M, POBLETE B. Meta-level sentiment models for big social data analysis[J]. Knowledge-Based Systems, 2014, 69(10): 86-99.

[4] PENALVER M I, GARCIA S F, VALENCIA G R, et al. Feature-based opinion mining through ontologies[J]. Expert Systems with Applications, 2014, 41(13): 5995-6008.

[5] GAUTAM G, YADAV D. Sentiment analysis of twitter data using mach-ine learning approaches and semantic analysis[C]// Proceeding of 7th International Conference on Contemporary

Computing(IC3). 2014:437-442.

[6] GARG Y,CHATTERJEE N. Sentiment Analysis of Twitter Feeds[C]// Proceeding of the 3rd Springer International Conference on Big Data Analytics. 2014:33-52.

[7] VANZO A,CROCE D,BASILI R. A context-based model for Sentiment Analysis in Twitter[C]// Proceeding of 25th International Conference on Computational Linguistics. 2014: 2345-2354.

[8] SONG S Y,LI Q D,LU D Y. Hot Event Sentiment Analysis Method in Micro-blog[J]. Computer Science, 2014, 39(6A): 226-260. (in Chinese)
宋双永，李秋丹，路冬媛. 面向微博客的热点事件情感分析方法[J]. 计算机科学，2014，39(6A)：226-260.

[9] LI Y Q,LI X,HAN X,et al. A Bilingual Lexicon-Based Multi-class Semantic Orientation Analysis for Microblogs[J]. Acta Electronica Sinica,2016,44(9):2068-2073. (in Chinese)
栗雨晴，礼欣，韩煦，等. 基于双语词典的微博多类情感分析方法[J]. 电子学报，2016，44(9)：2068-2073.

[10] LIANG J,CAI H M,YUAN H B,et al. Deep Learning for Chinese Micro-blog Sentiment Analysis [J]. Journal of Chinese Information Processing,2014,28(5):155-161. (in Chinese)
梁军，柴红梅，原慧斌，等. 基于深度学习的微博情感分析[J]. 中文信息学报，2014，28(5)：155-161.

[11] ZHANG L,QIAN G Q,FAN W G,et al. Sentiment Analysis Based on Light Reviews [J]. Journal of Software,2014,25(12): 2790-2807. (in Chinese)
张林，钱冠群，樊卫国，等. 轻型评论的情感分析研究[J]. 软件学报，2014，25(12)：2790-2807.

[12] LIU Z M,LIU L. Empirical Study of Sentiment Classification for Chinese Microblog Based on Machine Learning [J]. Computer Engineering and Applications,2012,48(1):1-4. (in Chinese)
刘志明，刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用，2012，48(1)：1-4.

[13] KWAK H,LEE C,PARK H,et al. What is Twitter,a social network or a news media? [C]// Proceedings of the 19th International Conference on World Wide Web(ACM). 2010: 591-600.

[14] VOLPENTESTA A P, FELICETTI A M. Identifying Opinion Leaders in Time-Dependent Commercial Social Networks[C]// Proceedings of 13th IFIP Working Conference on Virtual Enterprises(PRO-VE). 2012:571-581.

[15] MOMTAZ N J,AGHAIE A,ALIZADEH S. Identifying Opinion Leaders for Marketing by Analyzing Online Social Networks [J]. International Journal of Virtual Communities & Social Networking,2011,3(1):43-59.

[16] CAO J X,CHEN G J,WU J L,et al. Multi-Feature Based Opinion Leader Mining in Social Networks[J]. Acta Electronica Sinica,2016,44(4):898-905. (in Chinese)
曹玖新，陈高君，吴江林，等. 基于多维特征分析的社交网络意见领袖挖掘[J]. 电子学报，2016，44(4)：898-905.

[17] ZHANG W Z,HE H,CAO B. Identifying and Evaluating the Internet Opinion Leader Community Based on k-clique Clustering [J]. Neural Computing and Applications, 2014, 25(3/4): 595-602.

[18] XIAO Y,XU W,XIA L. Networking Groups Opinion Leader Identification Algorithms Based on Sentiment Analysis [J]. Computer Science,2012,39(2):34-37. (in Chinese)
肖宇，许炜，夏霖. 一种基于情感倾向分析的网络团体意见领袖识别算法[J]. 计算机科学，2012，39(2)：34-37.

[19] MA N,LIU Y. Superedge Rank algorithm and its application in identifying opinion leader of online public opinion supernetwork [J]. Expert Systems with Applications,2014,41(4):1357-1368.

[20] FAN X H,ZHAO J,FANG B X,et al. Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader [J]. Chinese Journal of Computers, 2013, 36(2): 360-367. (in Chinese)
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.

[21] CHEN B,TANG X Y,YU L,et al. Identifying Method for Opinion Leaders in Social Network Based on Competency Model [J]. Journal on Communications,2014,35(11):12-22. (in Chinese)
陈波，唐相艳，于泠，等. 基于胜任力模型的社交网络意见领袖识别方法[J]. 通信学报，2014，35(11)：12-22.

[22] WANG C X,GUAN X H,QIN T,et al. Modeling on Opinion Leader's Influence in Microblog Message Propagation and Its Application [J]. Journal of Software,2015,26(6):1473-1485. (in Chinese)
王晨旭，管晓宏，秦涛，等. 微博消息传播中意见领袖影响力建模研究[J]. 软件学报，2015，26(6)：1473-1485.

[23] ZHAI Z,XU H,LI J,et al. Sentiment Classification for Chinese Reviews Based on Key Substring Features[C]// Proceedings of 4th the Conference on Natural Language Processing and Knowledge Engineering. IEEE Computer Society,2009:1-8.

[24] YE Q,LIN B,LI Y J. Sentiment Classification for Chinese Reviews:A Comparison between SVM and Semantic Approaches [C]// Proceedings of the 4th International Conference on Machine Learning and Cybernetics. 2005:2341-2346.

[25] LIU H J,LI S J. Apply Fuzzy Comprehensive Evaluation to Establish Smartphone Assessment Model [J]. Computer Engineering and Applications,2016,52(1):224-228. (in Chinese)
刘焕军，李石君. 应用模糊综合评价进行智能手机评估建模[J]. 计算机工程与应用，2016，52(1)：224-228.

[26] YIN Y T,LI X M,CAI M S. 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.