文章编号：2095-2163(2020)11-0169-05 中图分类号：TP391.1 文献标志码：A

考虑情感强度的加权社会网络偏好信息识别研究

来能烨
（上海工程技术大学 管理学院，上海 201620）

摘 要：为了准确的识别网络文本中用户的情感偏好信息，提出了一种考虑情感强度的加权社会网络偏好信息识别算法。首先提取出文本特征项的独立信息，使用词典与互信息相结合的分词方法对文本做分词处理。分析实际文本中副词可以表达出的情感强度，将不同情感强度的副词赋予不同权重值，通过将句子本身定义的权重值与句中副词权值相乘来获得文本总情感强度。对提取出的特征项做向量转化处理，通过 GMM 算法进行情感偏好状态测定，完成识别全过程。仿真分析实验表明，本文算法可行性较强，识别效果较优。

关键词：加权社会网络；情感强度；特征提取；向量转化；情感偏好测定

Weighted social network preference information recognition considering emotional intensity

LAI Nengye
（School of Management, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】 In order to accurately identify the user´s emotional preference information in the network text, a weighted social network preference information recognition considering emotional intensity is proposed. First, we extract the independent information of the text features, and use the word segmentation method which combines dictionary and mutual information to segment the text. Analyze the emotional intensity that can be expressed by adverbs in the actual text, assign adverbs with different emotional intensity with different weight values, and obtain the total emotional intensity of the text by multiplying the weight value defined by the sentence itself and the weight value of adverbs in the sentence. Finally, the extracted feature items are transformed into vectors, and GMM algorithm is used to measure the emotion preference state to complete the whole process of recognition. The simulation results show that the algorithm is feasible and the recognition effect is better.

【Key words】 weighted social network; emotion intensity; feature extraction; vector transformation; emotion preference measurement

0 引 言

通常情况下，大多数网民会在各种社交网络上表达出对社会热点事件的不同看法。因此，如何有效识别其态度、行为和情感偏好程度成为被广泛关注的研究热点。情感偏好是情感强度的另一种表达形式，在根本上决定着人的思想、行为和生理活动，制约着情感的动力特性。

基于此，诸多学者及专业人士从各方面对该问题进行了研究并提出各自观点。Pablo C 等人[1]指出，社交网络领域的主要研究方向之一，是寻找和分析用户之间可能存在的联系。这些发展允许用户在其联系人网络上进行扩展，而不必在全部用户中进行搜索；Çavdar, A B 等[2]研究挖掘社交和交互数据，将这些信息与当前的数据分析模型结合起来，得出其结合程度是有限的的结论。可使用客户的社交网络信息来增强这个基本模型，以包含客户所做的间接贡献；Daniela F E 等[3]描述了如何将 Twitter 上的性别识别作为一种智能的商业工具，来确定用户之间的隐私问题，并最终为更有可能积极响应目标广告的客户提供更个性化的服务；Ran X[4]研究网络传播效应，也被称为同伴效应或社会影响过程，并提出了几种替代估计方法，当存在共同决定影响和选择的未观察特征时，这些方法有可能正确识别传染效应。采用蒙特卡罗模拟结果，设计了一种网络空间调整估计器；杜永萍等人[5]提出了一种 CNN-LSTM 模型下短文本情感分类方法，该方法以卷积神经网络模型为基础，构建大小不同的卷积窗口，对文本的谱义特征进行提取，采用长短时记忆模型，预测文本的情感倾向。通过在不同文本中进行验证，证明方法有效提高了网络文本情感识别的召回率，但是其准确率相对较低。穆永利等人[6]提出了一种基于 E-CNN 的情绪原因识别方法。该方法首先对本文进行卷积、池化等操作来融合句子中的语义信息，通过 CNN 集成降低数据不平衡性对识别效果的影响，解决了传统识别方法规则制定繁琐、需要对文本进行空间降维等问题。该方法可以从所有信息

作者简介：来能烨(1994-)，女，硕士研究生，主要研究方向：网络舆情管理。

收稿日期：2020-08-26

中有效识别全局信息,但是没有给出一个能够判断句子中真正情感的子句的合理度量,使得最终识别结果不够准确。

为提高情感偏好识别的准确率和识别效率,本文提出了一种考虑情感强度的加权社会网络偏好信息识别算法。该算法的优越之处在于将网络文本语句中不同程度的副词赋予不同的权重值,通过本身定义的权重值与句中的副词权值相乘来获得文本的总体情感强度。通过 GMM 算法进行情感偏好状态测定,完成识别全过程,总体识别效果更好,具有较好的应用价值。

1 文本挖掘

文本挖掘以语言学、统计梳理分析等作为主要理论依据,在信息检索技术的基础上,从网络繁杂的用户信息中,将能够表现出各类特征的独立信息提取出来。在文本挖掘过程中,文本分词是很重要的一部分,其关键部分在于歧义切分。在英文文本中,因其单词之间有空格能够被视为分隔符,所以歧义切分过程较为方便,但是中文文本中每句话的字词都是相互联系的,没有明显的分隔标记,相对英文文本来说,中文文本的歧义切分较为复杂。

为了使分词具有较好效率的同时也能充分保证分词的准确性,使用词典与互信息相结合的分词方法,对文本进行分词处理。将 $MI(x,y)$ 定义为词 x 和词 y 的互信息,则有:

$$MI(x,y)=\log\frac{P(x,y)}{P(x)P(y)},\ P(x,y)=\frac{c(x,y)}{\sum_{x',y'}c(x',y')}. \quad (1)$$

式中,当 $MI(x,y)\geqslant 0$ 时,表明二者经常同时出现,同时证明两个词的关联性很强;当 $MI(x,y)\approx 0$,则代表 x 和 y 同时出现的次数极少,从而证明二者的关联性较弱;当 $MI(x,y)\leqslant 0$ 时,则表明 x 和 y 不会同时出现,二者之间没有关联性,为互补分布。

通过对词语互信息的计算,原词典中信息就会随之丰富,从而获得词与词之间的互信息矩阵为:

$$\begin{bmatrix} 1 & MI(A,B) & MI(A,C) & \cdots \\ MI(A,B) & 1 & MI(B,C) & \cdots \\ MI(A,C) & MI(B,C) & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (2)$$

在进行文本分词时,为丰富词典信息,使用双向匹配分词法对网络文本语句进行切分处理。在处理过程中,当正向和逆向切分的最终呈现效果不同时,通过互信息选出最适合整体的分词结果,同时计算切分后词语的整体平均互信息以减少词语个数对切分结果的影响。其计算方法如式(3):

$$E=\frac{\sum_{i=1,j>1}^{n}(W_i,W_j)}{n!}. \quad (3)$$

式中,n 表示被切分词语数量,W_i 表示第 i 个切分词语。

由于中文文本中的语言表达形式较为复杂,直接挖掘分析切分后的语句尤为困难。因此需要将分词处理的文本整合成更适合定量研究的文本情感形式。首先,提取各网络文本内的情感特征项,然后对提取后的情感特征项做文本系统结构化,并将其作为中间状态依次对文本信息进行描述。在文本系统中,文本之间是相互不发生联系的,因此从数据整体来看文件之间数据是没有结构关系的,而结构化就是将程序中逐渐积累出的内容和数据进行归纳整理,使程序数据条理化,更易于后期的处理。

文本通常能够通过词语来表达特征,如关键词、主题词、短语等。一般情况下,文本特征大致可以划分为语义特征和描述特征两类,通过处理特征项就可以实现文本分析。提取语义特征中的评价对象主要过程如下:

(1)采用中分词方法对文本进行分词处理。

(2)对切分后的名词进行比对,得到评价对象。

(3)选出文本中含有评价对象的句子。

(4)将修饰评价对象的词语进行筛选,将其视为情感词,并且将修饰情感词的副词定义为修饰词。

(5)记录情感词及修饰词的相对位置。

2 情感强度模型

2.1 基于情感强度的词表构建

在文本中,句中的一些副词往往可以表达出这个句子的情感强度,不同程度的副词赋予不同的权重值。整个句子的最终情感权值,可以通过自身定义的权值与句中的副词权值相乘而获得。

本文选择 219 个程度副词,根据其强度分为 5 个等级(W_1,W_2,W_3,W_4,W_5),分别赋予不同的权重值见表 1,构建的文本情感见表 2。

表 1 程度副词权重

Tab. 1 Weight of degree adverbs

程度副词	程度级别	权重值
超级,特别,十分,极其,更加,绝对,无与伦比等	极其	1.25
多,实在,尤为,很是,较为,愈加等	较	1.0
略,稍微,有点,或多或少,略微等	稍	0.75
不,弱,不怎么,微,不甚,不大等	欠	0.5

表 2　情感词表

Tab. 2　Sentiment word table

大类	小类	权重值
乐	快乐、高兴等	1.5
好	相信、尊敬等	1.25
恶	嫉妒、怀疑等	1.0
哀	悲伤、失望等	0.75

2.2　文本情感计算规则

将文本 D 分解成句子 S 的集合，则 $D=\{S_1,S_2,\cdots,S_n\}$，每个句子的情感权值(S_i) 为：

$$F(S_i)=\sum S_{w_i}. \tag{4}$$

则整篇文本的情感权值为：

$$F(S)=\sum F(S_i). \tag{5}$$

式中，S_{w_i} 表示每个句子中副词的权重值；如果 $F(S)>0$，则可以判定该文本为正向情感；如果 $F(S)<0$，则可以判定该文本为负向情感；如果 $F(S)=0$，则可以判定该文本为中性情感。

计算情感词 W 的值 S_{w_i} 如下式：

$$S_{w_i}=P_{w_i}-N_{w_i}. \tag{6}$$

其中，P_{w_i} 和 N_{w_i} 的计算过程可用下式表示：

$$\begin{aligned} P_{w_i}&=\frac{fp_{w_i}}{(fp_{w_i}+fn_{w_i})}*\frac{N_p}{(N_p+N_n)},\\ N_{w_i}&=\frac{fn_{w_i}}{(fp_{w_i}+fn_{w_i})}*\frac{N_p}{(N_p+N_n)}. \end{aligned} \tag{7}$$

式中，N_p 表示正向词的数目，N_n 表示负向的词汇数目。

考虑到文本中句型对情感强度判定的影响，根据不同句型归纳出句子的情感值如下：

疑问句：$F'(S_i)=F(S_i)\times(-0.2)+(-0.5)$

反问句：$F'(S_i)=F(S_i)\times(-0.6)+(-0.5)$

感叹句：$F'(S_i)=F(S_i)\times(1.5)$

假设句：$F'(S_i)=F(S_i)\times(-0.2)$

通过句子的情感值可以获得文本的情感权重值为[7]：

$$F'(S)=\sum F'(S_i). \tag{8}$$

当 $F'(S)>0$ 时，则表示为正向情感[8-9]，$F'(S)<0$ 时，则可以定义为负向情感，$F'(S)=0$ 时，则文本可以定义为中性情感。

再次加入程度副词进行计算如下：

$$S_{w_i}=(P_{w_i}-N_{w_i})*(1\pm\sigma)*N_e. \tag{9}$$

式中，N_e 为否定系数，$*$ 为调节过程。

若感情词与否定词相邻，则可以判断该文本为负偏好情感，因此将其否定系数 N_e 设置为 -1。σ 表示调节系数，如果筛选出的情感词与程度副词“非常”、“极其”等相邻时，则可以判定其为正偏好情感，其表达式如下：

$$S_{w_i}=(P_{w_i}-N_{w_i})*(1+\sigma)*N_e. \tag{10}$$

如果情感词与“一般”、“还可以”等程度副词相邻时[10]，则可以将该文本定义为中偏好情感。则其情感得分如下式：

$$S_{w_i}=(P_{w_i}-N_{w_i})*(1-\sigma)*N_e. \tag{11}$$

3　加权网络信息偏好识别算法

通过计算用户对目标个体的情感偏好指数，可以了解用户对任意事物的选择倾向，并能反映出相对于他人的不同价值取向，即价值取向表现的优劣程度足以直接反映出个人的情绪偏好。

语篇情感偏好识别主要是通过语篇中句子的情感权重来判断。考虑情感强度的社会网络偏好信息加权识别，是在文本挖掘和情感强度模型建立的基础上，通过 GMM 算法进行特征提取和向量转换，确定情感偏好状态，完成识别过程。具体流程如图 1 所示。

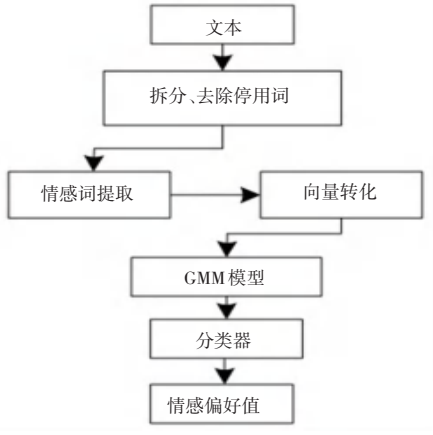

图 1　识别流程

Fig. 1　Identification flow chart

为了更有效的识别文本情感偏好，需对文本进行预处理。文本处理包括：命名实体及过滤停用词等。本文在 Windows 操作系统下，获取相关文本数据，并对文本中表情符号、网址等无意义的文本进行清理。

采用 GMM 算法识别情感词。其具体数学表达形式如下：

$$P(x_t\mid\lambda)=\sum_{i=1}^{m}a_ip_i(x;\mu_i,\sum i). \tag{12}$$

式中，x_t 为第 t 个高斯分布的 D 维随机向量[11]，a_i 代表第 i 个单高斯分布的权重值，且定义 $\sum_{i=1}^{m}a_i=1$ $p_i(x_i)(i,\cdots,m)$ 为高斯分布函数，则：

$$p_i(x_i;\mu_i,\sum_i)=\frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}}\times$$
$$\exp\{-\frac{1}{2}(x_t-\mu_i)^{\mathrm{T}}\sum^{-1}(x_t-\mu_i)\} \quad (13)$$

式中，$\sum_i$ 表示协方差矩阵，μ_i 表示均值矢量。协方差矩阵可以用满矩阵，也可以使用简化后的对角矩阵。高斯分布密度如下式：

$$\lambda_i=\{a_i,\mu_i,\sum_i\},i=1,\cdots,m. \quad (14)$$

式中，为了能够得到最佳的样本分布概率，采用 EM 算法来估计 GMM 模型的参数[12]。

设待测样本为 y，将分类器给出的似然度标记为 $P(y|\lambda_k)$，其中，k 代表各情感强度，则各情感强度权值如下：

$$I_n=\frac{\sum_{1\leqslant i\leqslant 4}|In(P(y|\lambda_k))InP(y|\lambda_k)}{\left|\sum^{4}In(P(y|\lambda_k))\right|}. \quad (15)$$

似然度直接决定分类器的置信度，更直接的表现是似然度的分散程度。置信度越高，则判定结果越准确。完成识别全过程步骤如下：

(1)将文本输入分类器，做词法和语法分析，获得更易识别的文本结构。

(2)对获得的结构化文本进一步分析，将其与相应的情感规则进行匹配。结合情感强度模型，做情感划分，得到情感值。

(3)输出情感值。将判断用户偏好的情感值反馈给机器。

(4)抽取反馈中有价值的信息，更新词典。

4 仿真实验

为了验证考虑情感强度的网络评论情感偏好识别方法的有效性，本文使用了八爪鱼采集器，爬取了新浪微博上关于“新冠肺炎疫情”爆发期间的热门评论，共计2 943条作为数据来源进行对比实验。

实验所用情感词主要来源于《知网》的情感分析用语词集，并且加入了一些最新出现的网络情感用词，对词语进行去重处理后，获得的主要情感词。

为验证本文算法的准确性，将文献[4-6]中提出的方法与本文算法进行比较。利用各算法的准确率 *Pre*、召回率 *Rec* 和 *F* 值作为评判项。*Pre* 其表达式为：

$$Pre=\frac{TP}{TP+FP}. \quad (16)$$

Rec 能够衡量系统查全率，其表达式为：

$$Rec=\frac{TP}{TP+FN}. \quad (17)$$

在识别过程中，往往不能够使准确度和召回率同时具有较好的表现，因此常使用 *F* 值来对识别的整体效果做评估。*F* 值的常用表达式如下：

$$F\text{ 值}=\frac{2\times Pre\times \mathrm{Rec}}{Pre+Rec}. \quad (18)$$

其中，各参数含义见表 3。

表 3 分类评价标准参数含义表

Tab. 3 Meanings of parameters for classification evaluation criteria

	分类为 T 类	分类为非 T 类
实际为 T 类	*TP*	*FP*
实际为非 T 类	*FN*	*TN*

各算法的各项指标值如下图 2 所示。

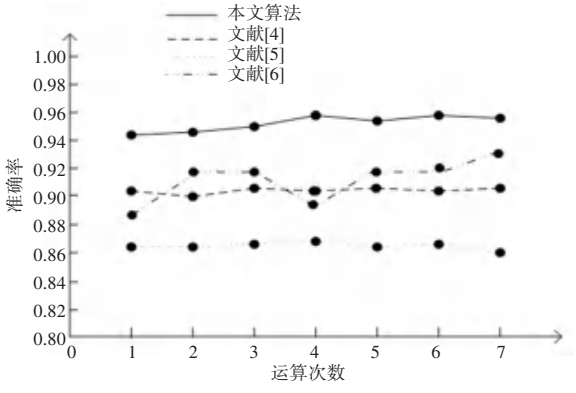

图 2 不同方法的准确率对比图

Fig. 2 Comparison of precision of each method

由图 2 可见，在 7 次迭代下，本文方法对网络偏好数据的分析与识别准确率较高，说明在进行语篇情感偏好识别时，对语篇中句子的情感权重判断效果较好。在文本挖掘和情感强度模型建立的基础上，考虑情感强度的社会网络偏好信息加权识别方法实际应用效果较强。

应用情感分析用语词集，在系统查全率即召回率方面进行对比结果如图 3 所示。

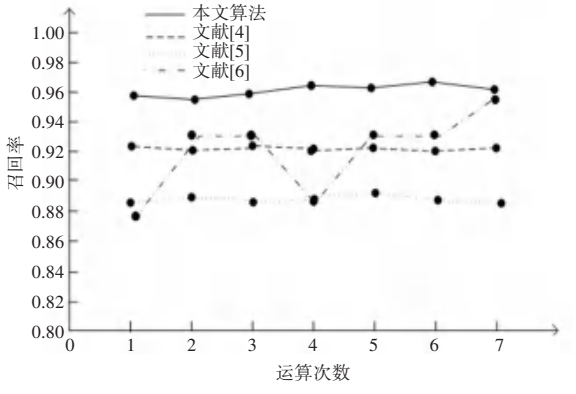

图 3 不同方法的系统数据召回率对比图

Fig. 3 Comparison of recall of each method

由图 3 可知，在进行系统召回率测试时，本文方法的召回效果对比结果鲁棒性较强，说明本文方法对文本中表情符号、网址等无意义的文本进行清理后，实际有用的数据能够被系统查全即有效召回。

将以上两次实验数据进行二次拟合，使用 F 值进行整体效果评估。评估结果如图 4 所示。

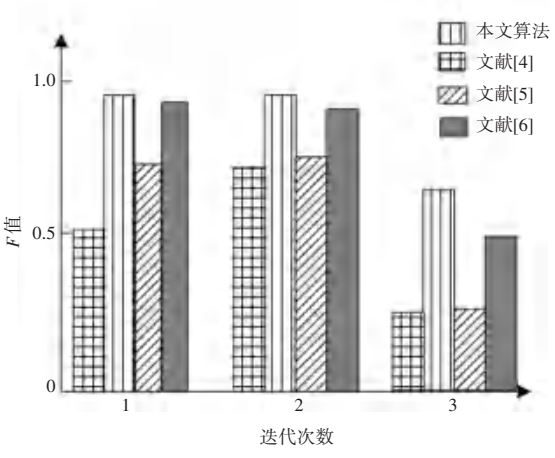

图 4　不同方法的 *F* 值对比结果

Fig. 4　Comparison of F values of each method

从图 4 中可以看出，在情感识别的过程中，文献[4]算法整体效果最差，本文算法要优于其它文献方法。最终获得的召回率、准确率和 *F* 值 的数值都高于前两者。因此，证明本文算法是可行的，并且识别效果更优。

5　结束语

用户对网络使用体验感要求增高，情感强度能够有效获取用户对某种属性喜爱程度。本文提出的考虑情感强度的加权社会网络偏好信息识别算法，经对比试验得出如下结论：

（1）通过将其本身定义的权重值与句中的副词权值相乘来获得文本的总体情感强度，优化语句情感权重，实现语句的整体阈值。

（2）在排除无意义文本信息后，对文本进行特征提取及向量转化，通过 GMM 算法进行情感偏好状态测定，总体识别效果更好。

参考文献

[1] PABLO C, ALBERTO R, SARA R, et al. Relationship recommender system in a business and employment-oriented social network[J]. Information Sciences, 2018, 433(34):204-220.

[2] ÇAVDAR, A B, FERHATOSMANOGLU N. Airline customer lifetime value estimation using data analytics supported by social network information[J]. 2018, 67(46):19-33.

[3] DANIELA F E, LU X. Twitter Users' Privacy Concerns: What do Their Accounts' First Names Tell Us? [J]. Nephron Clinical Practice, 2018, 3(1):40-53.

[4] RAN X. Alternative estimation methods for identifying contagion effects in dynamic social networks: A latent - space adjusted approach[J]. Social Networks, 2018, 54(12):101-117.

[5] 杜永萍, 赵晓铮, 裴兵兵. 基于 CNN-LSTM 模型的短文本情感分类[J]. 北京工业大学学报, 2019, 45(7):662-670.

[6] 慕永利, 李旸, 王素格. 基于 E-CNN 的情绪原因识别方法[J]. 中文信息学报, 2018, 32(2):120-128.

[7] 廖祥文, 谢媛媛, 魏晶晶, 等. 基于卷积记忆网络的视角级微博情感分类[J]. 模式识别与人工智能, 2018, 31(3):219-229.

[8] 王鹤琴, 王杨. 基于情感倾向和 SVM 混合极短文本分类模型[J]. 科技通报, 2018, 34(8):149-154.

[9] 卢新元, 卢泉, 黄梦梅, 等. 基于情感倾向的众包模式下接包方声誉评价模型构建[J]. 统计与决策, 2018, 34(17):177-180.

[10] 朱婉菁. 网络事件中公众行为偏好与政府网络治理策略的逻辑互动[J]. 天津行政学院学报, 2019, 21(6):35-42.

[11] 石雪, 李玉, 李晓丽, 等. 融入邻域作用的高斯混合分割模型及简化求解[J]. 中国图象图形学报, 2019, 22(12):1758-1768.

[12] 李荟, 赵云敏. GMM-UBM 和 SVM 在说话人识别中的应用[J]. 计算机系统应用, 2018, 27(1):225-230.

（上接第 168 页）

参考文献

[1] NIKAZOBEC. 上海城市垃圾管理研究[D]. 上海外国语大学, 2019.

[2] 熊孟清. 便宜行事，垃圾分类方能致远[N]. 中国环境报，2020-11-11(3).

[3] SAATY T L. How to make a decision: The analytic hier- archy process[J]. European Journal of Operational Re- search，1990，48(1):9-26.

[4] KARAGIANNIDIS A, MOUSSIOPOULOS N. Application of ELECTRE III for the integrated management of municipal solid wastes in the Greater Athens Area[J]. European Jour- nal of Operational Research，1997，97(3):439-449.

[5] 冯思静，马云东. 我国城市垃圾分类收集的经济效益分析[J]. 江苏环境科技，2006(1):49-50.

[6] 邢建平. 运输问题表上作业法改进思路研究论述[J]. 科技创新与应用，2015(35):57.

[7] 王锟，李小琴. 基于表上作业法对物资转运问题的探讨[J]. 科教导刊(上旬刊)，2016(7):161-163.

[8] 曾强，沈玲. 产销平衡运输问题表上作业法的计算机辅助教学方法[J]. 教育教学论坛，2018(42):259-260.

[9] 高岩. 运筹学在选址问题中的应用[J]. 内江科技，2014，35(7):47-48，58.

[10] 鲁晓春，詹荷生. 关于配送中心重心法选址的研究[J]. 北方交通大学学报，2000(6):108-110.

[11] 刘会茹，张红梅. 运筹学在现代物流中的应用[J]. 科技视界，2015(34):159-160.