基于 HMM 的个体微博情感分析预测方法研究*

郭福亮　周　钢
（海军工程大学电子工程学院计算机工程系　武汉　430033）

摘　要　微博已经逐渐成为个体表达情感的重要途径，进行个体情感分析预测研究具有重要意义。在研究个体情感特点基础上，设计了基于 HMM 的个体微博情感分析模型，通过微博文本情感提取，PAD 情感模型维度量化和 HMM 模型对个体情感发展进行分析步骤获取个体情感状态变化稳定特征，并进行情感预测以实现及时心理疏导，通过实例研究发现该方法在无外部重大事件刺激下具有很好预测效果。

关键词　HMM；情感分析；PAD 模型；微博
中图分类号　TP393　　**DOI**：10.3969/j.issn1672-9730.2014.02.027

Micro-blog Individual Emotional Analysis and Forecast Method Based on HMM Model

GUO Fuliang　ZHOU Gang
(College of Electronics Engineering, Naval University of Engineering, Wuhan　430033)

Abstract　Microblogging has gradually become an important way to express individual emotion, the study on the analysis and forecasting of individual emotional has important significance. On the base of studying Emotional characteristics of individuals, the HMM-based model of the individual micro-blog sentiment analysisbased is proposed. The model achieves in accordance with the three-step: the microblogging text sentiment extraction, quantification through the PAD Dimensional model, and forecasting the development of the individual emotion based on HMM model. The purpose of the model is to analyze the stable characteristics of the individual emotional state and make emotional predicted to achieve timely psychological counseling. Through case studies found that the method has a good prediction in the absence of external stimuli major event.

Key Words　HMM, sentiment analysis, PAD model, micro-blog
Class Number

1　引言

人类的情感是一种多成分、多维量、多种类、多水平整合的复合心理过程，是对趋向知觉为有益的，离开知觉有害的事物的体验倾向[1]。对于人类情感的研究主要集中在情感的产生、识别和表达三个方向，语言是个体情感外在表现的重要途径之一，分析个体语言表达是进行情感识别研究的重要方法。2010 年中国互联网舆情报告指出，微博成为网络舆论主要载体[2]，大量微博用户发布的文本信息表达个体情绪情感，通过对用户文本情感提取、识别个体情感，对预防个体极端情绪和行为倾向有重要参考意义。

目前，微博情感分析主要针对大量微博文本情感倾向进行统计分析，研究在商品评价和舆情监控等领域的应用。本文则在研究个体情感发展特点基础上，分析当前文本情感提取基本方法，提出了基于隐形马尔科夫链方法的情感分析模型，从而分析微博用户情感及其发展倾向预测。

2　研究背景

2.1　情感特点

情感是一个十分复杂的现象，包含丰富的内

* 收稿日期：2013 年 8 月 21 日，修回日期：2013 年 9 月 30 日
作者简介：郭福亮，男，博士，教授，博士生导师，研究方向：分布式处理与网络技术、计算机应用技术。周钢，男，硕士，研究方向：分布式处理与网络技术。

容。情绪和情感是人对客观事物的态度体验,是人的生理需求和社会需求是否获得满足的反映,包括了情感过程和情感个性[3]两个层次:

1)情感过程:情感的具体表现为情绪和心情两种情感状态,情感状态变化是变化过程的一个重要方面。当为情绪状态时,激动水平较高,有强烈的情绪表现和明显的情感行为;当为心情状态时,激动水平较低,变化缓慢,没有强烈的情绪表现和明显的情感行为;从性质上情感状态可分为高兴、愤怒、恐惧、悲哀等。从变化方式上情感过程可分为受外界刺激影响的应急变化和由自身特性决定的自然变化。

2)情感个性:情感个性是与个体特性相关的包括需求、动机、兴趣等个性倾向性和能力、性格、气质等个性心理特征。情感态度表示对人或事物在态度方面的比较稳定的评价性情感,包括褒贬、喜恶等;情感在气质、性格方面主要体现情绪体验的强度、情绪状态变化的速度、情绪的稳定性和持久性,以及在同样外部刺激条件下产生某种情绪倾向性大小等。

情感具有多维度结构,情感的表示可看作具有信息度量的多维空间的点在情感空间中的映射[4]。情感映射维度论认为不同情感是逐渐的、平稳的转变,不同情感之间的相似性和差异性由维度空间距离显示。代表性的情感维度模型包括:1966 年 Watson 等设计的二维量表分析模型,设计了正负性情绪量表 PANAS 和症状自评量表 SCL-90 等,1974 年 Mehrabian 和 Russell 提出的 PAD 三维情感模型,认为情感具有愉悦度、激活度和优势度三个维度,以及 R. Plutchik 提出结合维度和基本情绪理论的情感锥球模型[5]。

2.2 情感分析模型

情感分析模型是采用数学方法分析人类情感,实现情感的模型化和形式化[6]。依照描述情感的数学方法可分为维度空间、非线性、灰色理论和随机过程等。

情感分析模型建模主要先对情感进行量化分析,采用情感维度或情感熵等方法,然后对情感自然个性和外界刺激等方面进行分类分层考虑,从概率转移、分层网络等方面研究情感状态变化,完成对情感的分析。情感分析模型研究主要方法包括基于欧氏空间的情感数学模型,建立情感空间的概率模型进行分析计算;基于马尔科夫链的情感计算模型,建立情感概率空间从而实现情感变化的模型模拟,给出了情感能量、情感强度和情感熵等概念;基于自组织理论的情感建模,依据情感由基本情感和表征人意志力的内驱力形成,借鉴自组织理论,模糊数学,最优化理论等数学思想,构建不同性格特征的数学模型;基于贝叶斯网络的情感分析建模,定义了性格空间和情感空间,设计了分层和网络化的情感分析模型[7]。

2.3 文本情感提取

文本情感分析主要任务就是根据文本来判断作者的情感倾向,主要利用底层情感信息抽取的结果将情感文本单元分为若干类别,如分为褒贬,喜悲等对立两类或更为细致的感情类别(如喜怒哀乐等),并进行分析归纳。文献[7]最早给出了情感分析的概念,文献[8]针对中文的文本情感分析的任务、内容和主要技术进行描述。

文本情感分析可分为三个研究层次,即情感信息的抽取、情感信息的分类以及情感信息的检索与归纳。其中情感信息抽取是抽取情感文本中有价值的情感信息,是情感分析的基础任务,为后续文本情感分析提供数据基础。

文本情感提取按照处理文本的粒度不同可以分为词语级,语句级和篇章级;按照不同分析目的,可以分为主客观分析和主观分析,前者主要研究作者对客观事物的褒贬评价,后者则主要研究作者自身的喜怒感受;按照分析内容的不同,可分为对新闻事件的情感分析和对商品评价的情感分析;按照技术处理手段可分为基于词典的情感分析和基于机器学习的情感分析,前者主要是利用基础情感词典对文本中词语进行情感分析,后者则是利用 SVM 方法、神经网络、朴素贝叶斯等分类器进行文本情感分析;按照有无人工参与可分为无监督分类方法和有监督分类方法,主要区别在于是否需要人工词语情感标注。

结合微博文本的长度较短,结构不规范,中文语法结构复杂等特点,本文的文本情感提取算法主要基于情感词典的方法。

3 基于改进 HMM 的情感分析模型

本文主要通过个人微博文本进行情感提取,建立适当模型研究个体情感发展趋势。隐形马尔科夫模型(Hidden Markov Model,HMM)是一种用参数表示,用于描述随机过程统计特性的概率模型,因具有成熟算法及其数据处理中表现很好的鲁棒性广泛应用于自然语言处理、文本分类等领域[9]。

3.1 隐马尔科夫模型

按照系统的发展,将时间离散化为事件节点,

对应的系统状态用随机变量表示为一定的发生概率，这个概率成为状态概率。当系统由随机过程中的某一个阶段状态转移到另一个阶段状态时，在这个转移过程中存在着转移的概率，称为转移概率。如果转移概率只和目前相邻两个状态的变化有关，也就是说下一阶段的状态只和现在状态有关而与过去无关，这种离散状态按照离散时间的随机转移系统过程，称为马尔科夫过程[10]。

HMM 是在马尔科夫链的基础上发展起来的。在实际问题中，由于观察值和状态值通常不是两相对应的，二者通过一定的概率分布描述，实质是一个“双重随机过程”。其中 T 为观察值的时间长度，马尔科夫链过程通过转移概率(π,A)描述状态之间的转移，确定状态序列，随机过程通过观察值概率矩阵 B 确定观察值和状态之间的对应关系得到对应的观察值序列。

隐马尔科夫模型的定义如下：$\lambda=\{X,O,\pi,A,B\}$，由五个部分组成，详细含义如下：

1）设 X 表示状态的集合，其中 $X=\{S_1,S_2,\cdots,S_N\}$，N 表示状态的个数。在 t 时刻的状态用 q_t 表示。虽然状态是隐藏的，但在很多应用中，物理意义和状态或者状态集合相关。状态之间的内部关系，即从一个状态转移到另一个状态。

2）O 用来表示一组被观察值的集合。$O=\{V_1,V_2,\cdots,V_M\}$，M 的含义是某状态可输出的不同观察值个数。

3）状态转移概率矩阵 $A=\{a_{ij}\}$，矩阵元素的含义是从一个状态转移到另一个状态的概率。$a_{ij}=P(q_{t+1}=S_j\mid q_t=S_i)$其中 $1\leq i,j\leq N$。某些情况下，若任意两个状态之间都可以一步达到，则 $a_{ij}\geq 0$，也就是说状态转移概率矩阵的元素值都大于 0。

4）状态 j 时的观察概率矩阵 $B=\{b_j(k)\}$是在状态为 j 的情况下，其相应观察值的概率求解方式为 $b_j(k)=P\{O_t=V_k\mid q_t=S_i\}$，其中 $1\leq j\leq N$，$1\leq k\leq M$。

5）初始状态 $\pi=\{\pi_1,\pi_2,\cdots,\pi_N\}$，其含义是在初始时刻为某个状态概率，其表达式 $\pi_i=P\{q_1=S_i\}$，其中 $1\leq i\leq N$。

HMM 模型的基本要素由五个部分组成，也可简写成 $\lambda=\{\pi,A,B\}$，前文提到 HMM 模型是双重随机过程，在表达式中也体现了这一点，三个关键元素实际上可以分为两个部分，用 π,A 来说明马尔科夫链，即根据初始值和状态数可画个有向图，观察概率矩阵 B 来描述随机过程。

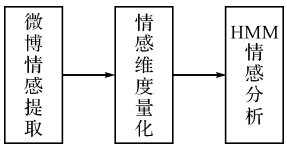

图 1　情感分析模型结构图

3.2　情感分析模型

本文建立针对个体微博文本的情感分析模型，模型基本结构如图 1 所示。

其基本步骤为

1）微博文本情感提取，采用基于情感词典方法对微博文本基本情感进行提取。

2）微博情感量化分析，采用情感维度对提取情感进行量化。

3）利用 HMM 对情感进行分析，完成情感状态转移概率研究，进行分析预测。

3.2.1　微博文本情感提取

本文的微博文本情感提取采用基于情感词典的方法，采用“知网”(Hownet)的语义词典，是一款为自然语言处理研究人员使用的一个共享软件。它是将汉语及英语词语所代表的概念作为描述对象，以展露概念之间、概念所包括的属性之间的关系作为基础内容的常识知识库。它所反映的内容包括概念的共性及个性，同时还展示了概念之间及概念的属性之间的各种关系。知网知识库内容包括中英双语知识词典、义原分类源文件、知网管理工具以及一些说明文件[11]。

本文采用基于知网的方法对微博文本中情感词以及否定副词和句法结构按文献[12]方法进行情感倾向判别，同时研究文本中副词、标点符号等采用文献[13]方法对情感倾向程度进行量化分析，同时采用文献[14]方法对评价的主客观进行分析判别。

3.2.2　情感维度量化

根据微博情感提取的情感倾向 T、强度 I 以及主客观情感类型 SI，采用 PAD 情感维度方法构建情感空间。PAD 三维情感模型将情感分为愉悦度、激活度和优势度，其中 P 代表愉悦度，表示个体情感状态的正负特征；A 代表激活度，表示个体的神经生理激活水平；D 代表优势度，表示个体对其他事物的控制状态。

采用统计学方法建立个体微博文本情感特征：倾向 T，强度 I 和主客观类型 SI，同 PAD 模型 P、A 和 D 之间的映射关系。研究不同个体已发布的 300 条带情感色彩的微博，通过基于知网的情感提取得到 T、I 和 SI 值，14 名专家针对各微博给出 PAD 值，按照 GEP 算法[15]得到映射公式：

$$\begin{cases}P=(T-T_{ave})/(T_{\max}-T_{\min})\\A=(T-T_{ave})I/|T-T_{ave}|\\D=(SI-SI_{ave})/(SI_{\max}-SI_{\min})\end{cases}$$

部分基本情感状态与 PAD 模型空间存在对应关系[16]如表 1 所示。

表 1 情感状态空间与 PAD 空间对应表

情感状态	P 值	A 值	D 值
平静	0.00	0.10	0.05
高兴	0.40	0.20	0.15
愤怒	−0.51	0.59	0.25
恐惧	−0.64	0.60	−0.43
悲伤	−0.40	−0.20	−0.50
厌恶	−0.40	0.20	0.10

3.2.3 基于 HMM 情感分析

设计基于 HMM 的情感分析模型,建立简单的六种情感状态的集合 $X=\{X_i\}$,$X_i=\{$平静,高兴,愤怒,恐惧,悲伤,厌恶$\}$,微博文本的观测集合 $O=\{O_i\}$,$O_i=\{$倾向 T_i,强度 I_i,主客观 $SI_i\}$,对微博文本按发布时间分为时间序列 $1,2,\cdots,i,N$,通过对应映射关系和隶属函数概率得到与集合 X 的对应。

HMM 模型进行微博文本情感分析,在给定模型的情况下观察序列 O 的概率,如何快速地选择在一定意义下“最优”的状态序列,使得该状态序列“最好地解释”观察序列,以及可能的模型空间,如何来估计模型参数,也就是说,如何调节模型$\{\pi,A,B\}$的参数,使得 $P(O|\lambda)$最大。

按照前后向递推法,由模型 λ 得到观察序列 O 的概率:

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)$$

按照 Viterbi 算法,在给定观察序列 O 和模型 λ 的条件下 t 时刻处于状态 S 的概率:

$$\lambda_t(i)=\frac{a_t(i)\beta_t(i)}{P(O|\lambda)}=\frac{a_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$$

且有:$\sum_{i=1}^{N}\lambda_t(i)=1$

采用 Baum-Welch 算法用于情感模型的参数估计:

$$\begin{cases}\bar{\pi}_i=\sum_{l=1}^{L}\alpha_1^{*(l)}(i)\beta_1^{*(l)}(i)\\ \bar{a}_{ij}=\dfrac{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\alpha_1^{*(l)}(i)a_{ij}b_j(O_{t+1}^{(l)})\beta_{t+1}^{*(l)}(j)/\phi_{t+1}}{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\alpha_1^{*(l)}(i)\beta_1^{*(l)}(i)}\\ \bar{b}_{jk}=\dfrac{\sum_{l=1}^{L}\sum_{t=1,O_t=V_k}^{T_l}\alpha_1^{*(l)}(i)\beta_1^{*(l)}(j)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\alpha_1^{*(l)}(i)\beta_1^{*(l)}(j)}\end{cases}$$

其中,ϕ 为对向前变量 a 和向后变量 β 进行处理的比例因子,l 为对应观察序列的序号。

那么建立 HMM 情感分析模型如图 2 所示。

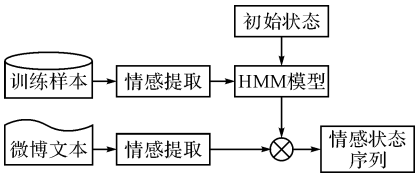

图 2 微博文本情感分析 HMM 模型框架

HMM 模型进行个体微博文本情感分析的基本步骤为:

1) 设置模型初始值:隐状态数 $L=6$,观察状态数 $N=3$,初始情感为平静 $\Lambda=(1,0,0,0,0)$。

2) 模型参数计算:使用训练样本按照上述方法得到转移概率矩阵 A 以及输出观察状态概率矩阵 B,从而得到 HMM 模型(A,B,Λ)。

3) 分析对象状态提取:按天进行时间序列划分,进行文本情感提取倾向 T_0,强度 I_0,主客观 SI_0,利用隶属函数得到初始状态概率 Λ_0。

4) 情感预测:根据初始状态 Λ_0 和 HMM 模型预测时间序列后的情感状态 O_{t+1}。

4 实例及分析

4.1 实验数据及指标

个体在日常生活中无重大外部时间刺激下,其情感发展和情绪变化具有相对稳定变化状态,其特点与个体本身特征相关。根据上述 HMM 模型方法可以分析得到微博用户个体的情感状态转移变化过程,从而能对个体心理情感进步发展进行预测并及时予以疏导和干预。

本文选用新浪微博中某活跃用户 P 连续发布的 200 篇微博,采用情感词典 Hownet 进行情感提取,按照映射公式得到 PAD 值,并根据隶属函数和情感对照表 1 得到微博的六种基本情感状态,按照 HMM 模型方法得到情感状态转移矩阵:

$$A=\begin{bmatrix}0.252&0.152&0.132&0.095&0.190&0.179\\0.125&0.196&0.113&0.156&0.202&0.208\\0.152&0.168&0.193&0.188&0.103&0.196\\0.102&0.177&0.183&0.206&0.162&0.170\\0.153&0.103&0.196&0.195&0.225&0.128\\0.216&0.204&0.183&0.160&0.095&0.175\end{bmatrix}$$

以及情感状态分布矩阵:

$$B=\begin{bmatrix}0.323&0.425&0.252\\0.256&0.281&0.463\\0.421&0.294&0.285\end{bmatrix}$$

(下转第 146 页)

[9] 王同标,刘念华. 正负折射率材料组成的一维光子晶体的能带及电场[J]. 物理学报,2007,56(10):5878-5882.
[10] Liang H, Chen-Ying Y, Wei-Dong S, et al. Design of incident angle-independent color filter based on sub-wavelength two-dimensional gratings[J]. Acta Phys. Sin,2012.

(上接第 102 页)

那么,矩阵 A,B 表征了用户 P 个体情感变化的固有稳定特征,结合某时刻 T 发布微博文本情感状态概率 $\Lambda_t=(1,0,0,0,0)$,按照 HMM($\Lambda_T,A,B,$)模型得到最大概率情感状态序列,从而得到 $T+t$ 时刻的用户 P 的情感状态。

4.2 实验结果

对于该个体 10 个阶段不同初始状态的 $t=10$ 天发布的微博文本进行情感分析,预测 10 天后情感特点并根据实际发布的微博情感进行比较研究。发现 10 个分析案例中其中 7 个能得到很好预测,其余 3 个预测结果与实际出入较大,发现均为外部重大事件发生导致。分析可以得到以下结论:

1) 在个体情感稳定情况下,模型能较好地预测分析个体情感发展。

2) 在外部重大事件突发时,模型矩阵 A,B 以及部分参数应当进行调整。

3) 模型具有较好的适应性,在多种初始状态情感和发展变化下,均能较好完成预测。

结合实例分析,对个体微博文本 HMM 情感分析模型下一步可以从以下几点进行改进研究:

1) 根据微博文本情感实时变化建立反馈参数,对模型进行调整。

2) 对进一步研究个体应对外部时间刺激的反映程度对其稳定情感模型影响,并针对不同类型时间建立不同影响模型,从而实现对外部事件刺激模型建立的完善。

5 结语

微博已经逐渐成为个体表达情感重要途径,研究微博情感和预测个体情感发展具有重要意义。本文建立了基于 HMM 的个体微博情感分析模型,通过微博文本情感提取,PAD 情感模型维度量化为 HMM 分析做好数据准备,采用 HMM 模型对个体情感发展进行分析得到个体情感状态变化稳定特征,并进行情感预测以实现及时心理疏导,通过实例研究发现该方法在无外部重大事件刺激下具有很好预测效果,下一步将针对外部事件刺激的情感变化进行研究,提高模型效能。

参 考 文 献

[1] Tracy J, Ramsey J. Emotions[M]. North Carolina: The Guilford Press,2001:21-25.
[2] 中国互联网信息中心. 第二十五次中国互联网发展状况统计报告[R]. 中国互联网统计报告,2010(1):1-10.
[3] 李维杰. 情感分析与认知[J]. 计算机科学,2010(7):11-16.
[4] Picard R W. Affective Compution[M]. London: MIT Press,1997:12-17.
[5] 王良志. 人工情感[M]. 北京:机械出版社,2009:39-49.
[6] 张颖,罗森林. 情感建模与情感识别[J]. 计算机工程与应用,2003(33):98-102.
[7] Bo Pang, Lillian lee. Thumbs up: Sentiment Classification Using Machine Learning Techniques [C]// EMNLP'02, July 6-7, Philadelphia, USA, 2002: 22-240.
[8] 魏韡,向阳,陈千. 中文文本情感分析综述[J]. 计算机应用,2011,31(12):3321-3323.
[9] 李开荣,孔照昆,陈桂香,等. 基于改进隐马尔科夫模型的文本分类研究[J]. 微电子学与计算机,2012(11):161-165.
[10] 李杰. 隐马尔科夫模型的研究及其在图像识别中的应用[D]. 北京:清华大学图书馆,2004:12-18.
[11] 董振东,董强,郝长伶. 知网的理论发现[J]. 中文信息学报,2007,21(4):3-9.
[12] 党蕾,张蕾. 一种基于知网的中文句子情感倾向判别方法[J]. 计算机应用研究,2010(4):1370-1372.
[13] 杨频,李涛,赵奎. 一种网络舆情的定量分析方法[J]. 计算机应用研究,2009(3):1066-1070.
[14] 蒙新泛,王厚峰. 主客观识别中的上下文因素的研究[J]. 中国计算机语言学研究前沿进展,2007-2009:594-599.
[15] Ferreira C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex System,2001,12(2):87-129.
[16] Gebhard P. ALMA-A Layered Model of Affect[C]// AAMAS'05. Utrecht, Netherlands: ACM, 2005: 29-36.