基于主题的文本句情感分析

王 磊 苗夺谦 张志飞 余 鹰
(同济大学计算机科学与技术系 上海 201804)
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 近年来,针对互联网在线信息的情感分析已经成为自然语言处理领域的一个研究热点。提出一个基于主题的情感向量空间模型,它将文本的潜在主题特征融入情感模型中,结合情感词典,利用多标签分类算法,对文本中句的情感极性进行分析与研究。实验结果表明,基于主题的情感向量空间模型在句的情感极性判断上取得了令人满意的效果。

关键词 情感词典,概率主题,多标签分类,情感分析
中图法分类号 TP391 **文献标识码** A

Emotion Analysis on Text Sentences Based on Topics

WANG Lei MIAO Duo-qian ZHANG Zhi-fei YU Ying
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)
(The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract The emotion analysis on internet online information has received much attention from natural language processing field in recent years. A novel emotion vector space model based on topics for text sentences was proposed. The new model including the latent topics features, emotion dictionary and multi-label classification algorithm was applied to analyze the polarity of sentences. Experiment result shows that the model is reasonable and effective in recognizing the polarity of sentences.

Keywords Emotion dictionary, Probability topics, Multi-label classification, Emotion analysis

1 引言

随着互联网的迅猛发展,"用户参与,用户体验"的思维理念不断深入人心,越来越多的普通用户乐于在互联网上抒发个人情感,评论产品性能,讨论时事政策,由此产生了大量带有个人主观情感特征的在线信息,如个人博客、产品评论、新闻评论等。这些信息都不同程度地反映了人们的各种喜好和情感倾向,如喜、怒、哀、乐等等。通过对在线信息的情感分析,可以很好地认识个体的情感状态,了解用户对产品的喜爱程度,但这一切仅仅依靠人工处理已经无法胜任,这就促进了情感分析技术的发展,使之成为自然语言处理领域的一个研究热点。

文本情感是情感分析领域的一个重要组成部分,依据其分析粒度的不同,可以分为相互依赖、相互支撑的3个层次[1]:词的情感分析、句的情感分析和篇章的情感分析。词的情感分析研究主要集中于如何识别情感词及情感词的极性,篇章的情感分析主要研究如何利用各种不同技术来识别篇章中的情感要素并判断篇章的情感极性。在3个层次之中,句的情感分析具有承上启下的作用,既依赖于词的情感特征分析,又能够获取更丰富的情感要素,为篇章的情感分析提供支持。

句的情感分析研究主要集中在对句子中的各种主观性信息的识别,判断句子的情感极性以及从句中提取出与情感极性分析相关联的各类要素,包括评价拥有者、评价对象、情感强度等。纵观以前的研究工作,对句的情感极性判断存在两种研究思路:基于情感知识的方法和基于特征分类的方法。基于情感知识的方法主要依靠已有的情感词典及带有情感标记的评价短语进行情感分析。该方法首先识别句中的评价词语或评价短语的情感极性,然后进行极性加权求和,从而识别出句子的情感极性。该方法的研究重点一般放在评价词语或评价短语的抽取和极性判断上。Hu和Liu[2]利用WordNet的同义词与反义词关系,识别情感词语的情感极性,并根据句中情感极性占优势的情感词对句的情感极性进行判断。基于特征分类的方法主要是采用机器学习的方法,利用语料库,选取大量有意义的特征来完成句的情感极性分类任务。Dave等[3]利用机器学习的方法,对句的情感极性识别进行了研究。

到稿日期:2013-05-21 返修日期:2013-08-24 本文受国家自然科学基金(61075056,61273304),中央高校基本科研业务费专项资金资助。
王 磊(1976—),男,博士生,CCF学生会员,主要研究方向为粗糙集理论、数据挖掘和机器学习等,E-mail:dragon_wlei@126.com;苗夺谦(1964—),男,教授,博士生导师,CCF高级会员,主要研究方向为粗糙集理论、粒计算、Web智能、模式识别等;张志飞(1986—),男,博士生,CCF学生会员,主要研究方向为粗糙集理论、粒计算、数据挖掘和机器学习等;余 鹰(1979—),女,博士生,CCF学生会员,主要研究方向为数据挖掘、粒计算等。

他们首先收集了1000多篇已经标注情感极性的评论文章，统计词语的n-gram组合在文档中的出现频率，并以出现频率的比例为基础对这些特征的情感极性进行评分，再以这些特征及其评分值来判断新句子的情感极性。

本文将句的情感极性分析作为研究重点，将潜在主题特征引入句的情感极性判断中，提出一个基于主题的文本句情感向量空间模型来对文本中句的情感极性进行多标签分类。实验结果表明，该模型取得了令人满意的效果。

本文第2节简要介绍了预备知识，包括Ren_CECps中文情感语料库、LDA模型；第3节详细阐述了本文提出的基于主题的文本句情感向量空间模型；第4节描述了实验过程及其结果分析；最后对全文的工作进行了总结。

2 预备知识

2.1 中文情感语料库

本文采用Ren_CECps中文情感语料库[7]作为实验对象，该语料库从中文网站爬取中文博客文章作为初始文本语料，共包含了1487篇中文博客文章，共计11255个段落、35096个句子、878164个词语。

在Ren_CECps中文情感语料库中，人工标示出所有文本中与情感表达相关联的语言信息，整个文本标注共分为3个层次：文本层、段落层、句子层。句子层的情感标注是整个中文情感语料标注的基础，其标注对象有：情感词与情感短语及其情感类别与强度标注、情感主及情感对象标注、修辞手法标注、句的情感类别及其强度标注等。句子层的上一层为段落层，段落层的标注对象有：段落主题词标注、段落中心句标注、段落情感类别及其强度标注。Ren_CECps中文情感语料库的最高层面为文本层，文本层的标注对象与段落层基本一致。

目前，对于情感类别的划分还没有统一的标准，各国学者对此有着不同的认识，或只是简单地将情感划分为褒义与贬义。在Ren_CECps中文情感语料库中，将所有情感分为8类最基本的情感类别，这8类基本情感类别分别是：惊讶(surprise)、悲伤(sorrow)、喜爱(love)、高兴(joy)、憎恨(hate)、期待(expect)、焦虑(anxiety)、生气(anger)。文本、段落与句子的情感类别及强度都被标注为一个8维情感向量，表示形式如下：

$$\vec{e}=(e^1,e^2,e^3,e^4,e^5,e^6,e^7,e^8) \tag{1}$$

其中，e^i 被标注为8类情感类别中的一个基本情感类别的情感强度，其取值范围为0.1到1.0。

2.2 隐含狄列克雷模型

LDA模型[8]由Blei等人在2003年提出，它是一个“文本-主题-词”的三层贝叶斯产生式模型，其特点是参数空间的规模与语料库大小无关，适合于处理大规模语料库。

在LDA模型中，语料库中的每一篇文档可表示为一些主题所构成的一个概率分布，而每个主题则又是若干个词所构成的一个概率分布。最初的LDA模型只针对各文档的主题概率分布引入一个超参数使其服从Dirichlet分布，随后Griffiths等针对各主题的词分布也引入一个超参数使其服从Dirichlet分布，从而得到一个完整的生成模型。对于语料库中的每篇文档，LDA模型的文档生成过程如下：

1. 对每一篇文档，从主题分布中抽取一个主题；
2. 从被抽取到的主题对应的单词分布中抽取一个单词；
3. 重复上述步骤直至遍历文档中的每一个单词。

其生成过程的图模型如图1所示。

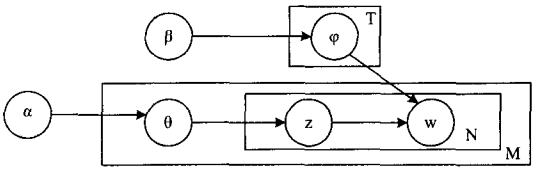

图1 LDA的图模型表示

3 基于主题的情感向量空间模型

3.1 情感词典

目前，有一些情感词典被开发出来用于情感分析，如英文情感词典WordNet-Affect和SentiWordNet，及HowNet所提供的VSA中英情感分析词汇。但至今为止，在中文情感分析领域仍缺少带有情感类别与情感强度的中文情感词典。本文中，情感分析研究完全基于Ren_CECps中文情感语料库。为了更好地完成情感分析研究，我们从该语料库中提取出所有带有情感极性的情感词及情感短语，构成一个全新的中文情感词典，并将它应用于情感分析实验中。

Ren_CECps中文情感语料库中，句子中的每一个情感词及情感短语都被标注为(惊讶，悲伤，喜爱，高兴，憎恨，期待，焦虑，生气)8个基本情感类别中的若干个，并标注了相应的情感强度，表示为如式(1)所示的一个情感向量$\vec{e}$。例如：情感词“无私”被标注为(0.0,0.0,0.7,0.0,0.0,0.5,0.0,0.0)，表达喜爱与期待，其情感强度分别为0.7与0.5。情感词“战争”被标注为(0.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0)，表达悲伤与憎恨，其情感强度分别为0.5与0.5。

在Ren_CECps中文情感语料库中，不同文章或不同句子中的相同情感词及情感短语具有不同的情感强度。从Ren_CECps中文情感语料库中提取所有的情感词及情感短语后，针对相同情感词及情感短语，计算它的相应情感类别的平均情感强度，并将这个平均值作为该情感词及情感短语的情感强度保存在情感词典中。

3.2 情感向量空间模型

根据“词袋”假设，将文本中的句子视为情感词及情感短语的组合，利用句中所包含的情感词及情感短语进行统计，可以识别句的情感极性。句的向量空间模型可以表示为：$\vec{S}=\{w_1,w_2,\cdots,w_n\}$，$n$ 为句中情感词及情感短语数目，w_i 为第 i 个情感词或情感短语。对于 w_i，赋给其一个情感向量$\vec{e}_i=(e_i^1,e_i^2,e_i^3,e_i^4,e_i^5,e_i^6,e_i^7,e_i^8)$，该句的情感向量空间模型可以进一步表示为如下形式：

$$\vec{S}=\{w_1,w_2,\cdots,w_n,\vec{e}_1,\vec{e}_2,\cdots,\vec{e}_n\} \tag{2}$$

3.3 基于主题的情感向量空间模型

针对一个文本中的所有句子，仅利用式(2)所示的情感向量空间模型来识别句子的情感极性，会割裂文中句子之间的相互关联，忽略了句子之间的上下文依赖关系。假设句子与文本描述的主题保持一致，那么句子与文本之间关于相同主题的情感描述也保持一致，则句子的情感极性应该由更能反映句子与文本主题特征的情感词或情感短语来决定。

本文将潜在主题特征融入句的情感向量空间模型，针对文档 D 引入LDA模型，得到 T 个隐含主题$\vec{T}=\{t_1,t_2,\cdots,t_T\}$以及主题-词的概率分布$\vec{\varphi}$，利用“文本-主题-词”之间的概率分

布来解决文本中句子之间的上下文依赖问题。作者提出一个基于主题的情感向量空间模型，从 T 个隐含主题中找出概率权重最大的主题 t_m，将句的情感向量空间模型进一步扩展，公式如下：

$$\vec{S}=\{w_1,w_2,\cdots,w_n,\vec{e}_1,\vec{e}_2,\cdots,\vec{e}_n,\varphi_{m1},\varphi_{m2},\cdots,\varphi_{mn}\} \quad (3)$$

w_i 表示第 i 个情感词或情感短语，$\vec{e}_i$ 表示为第 i 个情感词或情感短语的情感向量，φ_{mi} 表示基于主题 t_m 第 i 个情感词的主题-词的概率分布。

为了更加清晰地描述句子所具有的情感特征，对式(3)做进一步变换，得到基于主题的情感向量空间模型，其表示形式如下：

$$\vec{S}=(\sum_{i=1}^{n}(1+\varphi_{mi})e_i^1,\sum_{i=1}^{n}(1+\varphi_{mi})e_i^2,\cdots,\sum_{i=1}^{n}(1+\varphi_{mi})e_i^7,\sum_{i=1}^{n}(1+\varphi_{mi})e_i^8) \quad (4)$$

$\sum_{i=1}^{n}(1+\varphi_{mi})e_i^j$ 表示句中所有情感词或情感短语的第 j 个情感的情感强度加权求和，从而通过式(4)可以得到句子所具有的情感及情感强度。

3.4 句的情感极性判断基本框架

本文中，句的情感极性判断基本框架如图 2 所示，左侧是训练过程，右侧是测试过程，并将情感词典和 LDA 模型应用其中。句子的情感极性判断流程共分 6 步，具体描述如下：

步骤 1　从 Ren_CECps 中文情感语料库中随机抽取出实验数据，从而构成训练语料与测试语料；

步骤 2　利用实验语料，构造情感词典；

步骤 3　分别对训练语料与测试语料进行预处理，去除少量不含有情感极性的句子及句子中的非情感词与非情感短语，仅保留句中情感词与情感短语；

步骤 4　将情感词典与 LDA 模型分别作用于训练数据与测试数据，构成基于主题的情感向量空间模型；

步骤 5　对训练数据进行多标签分类训练，生成多标签分类模型；

步骤 6　将生成的多标签分类模型用于测试数据，并评价实验分类结果。

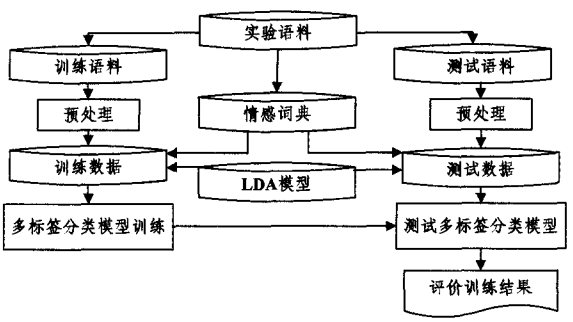

图 2　句的情感极性判断框架

4　实验与分析

4.1　实验数据

在实验中，从 Ren_CECps 中文情感语料库随机选取了 500 篇博客文章，共 12043 个句子。对数据集进行预处理：1)去除文本中少量的不含有任何情感极性的句子；2)去除句子中所有非情感词与非情感短语，仅保留情感词与情感短语。将预处理后的 500 篇博客文章平均分为 10 份，采用 10 折交叉验证。

4.2　实验设置

本实验目标是对文本中句的情感极性进行多标签分类。BR(Binary Relevance)方法是多标签分类算法的一个典型代表。BR 算法适用于标签数量 q 较小的情况，而本文采用的 Ren_CECps 中文情感语料库中情感类别标签只有 8 类，故采用 BR 算法作为多标签分类算法，并采用 Naïve Bayes 算法作为 BR 算法中的基本分类算法。

对实验结果的评价，本文采用基于标签的评价方法[10]。对于某个单个标签，采用公式 $M(tp_\lambda,tn_\lambda,fp_\lambda,fn_\lambda)$ 来评价单个标签 λ 的分类效果，tp_λ 表示正确的 positives 标签分类个数，tn_λ 表示正确的 negatives 标签分类个数，fp_λ 表示错误的 positives 标签分类个数，fn_λ 表示错误的 negatives 标签分类个数。多标签分类的宏平均和微平均公式如下：

$$M_{macro}=\frac{1}{|L|}\sum_{\lambda=1}^{|L|}M(tp_\lambda,fp_\lambda,tn_\lambda,fn_\lambda) \quad (5)$$

$$M_{micro}=M(\sum_{\lambda=1}^{|L|}tp_\lambda,\sum_{\lambda=1}^{|L|}fp_\lambda,\sum_{\lambda=1}^{|L|}tn_\lambda,\sum_{\lambda=1}^{|L|}fn_\lambda) \quad (6)$$

4.3　实验结果与分析

本文实验的情感类别为(惊讶，悲伤，喜爱，高兴，憎恨，期待，焦虑，生气)8 类，不考虑中性情感的句子。利用 LDA 模型来发现主题特征，参数设置如下：$\alpha=0.5$，$\beta=0.1$，$L=8$，主题数 $T=3$，以上参数均为实验经验值。

针对基于主题的情感向量空间模型，8 类基本情感类别单标签情感极性判断的正确率如图 3 所示。

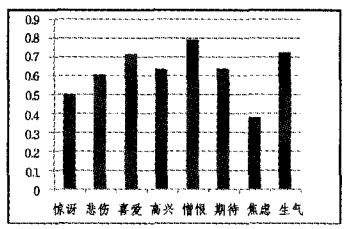

图 3　8 类基本情感类别单标签情感极性判断正确率

图 3 反映出惊讶与焦虑两类基本情感类别极性判断的正确率较低，这是由于 Ren_CECps 中文情感语料库中表达惊讶、焦虑两类基本情感类别的数据较少，对模型训练不够充分，从而影响了这两类数据在情感极性判断的正确率，今后需进一步充实、完善 Ren_CECps 中文情感语料库。

对比无主题特征的情感向量空间模型与基于主题的情感向量空间模型，在测试数据上进行文本句的情感极性多标签分类测试，表 1 分别给出两类模型的宏平均值与微平均值。

表 1　两类模型多标签分类结果的宏平均值与微平均值

	无主题特征模型	基于主题特征模型
宏平均精确率	0.782	0.803
宏平均正确率	0.607	0.626
微平均精确率	0.791	0.810
微平均正确率	0.612	0.631

表 1 显示出基于主题的文本句情感向量空间模型在句的情感极性判断上的优势。

结束语　本文重点对文本中句的情感极性问题进行了研究，提出了基于主题的情感向量空间模型。以 Ren_CECps 中文情感语料库为实验对象，利用抽取出的情感词典和 LDA 模型，解决了文本中句与句之间的上下文依赖问题，对语料库中文本句所具有的惊讶、悲伤、喜爱、高兴、憎恨、期待、焦虑、生

气 8 类情感类别进行多标签分类，取得了令人满意的结果。

文本只是利用主题特征来研究文本句的情感极性分类，目前的研究还有许多提高和改进的余地，如何以词和句的情感分析为基础研究篇章的情感问题也是今后进一步研究的方向。

参 考 文 献

[1] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848

[2] Hu Ming, Liu Bin. Mining and Summarizing Customer Reviews [C] // Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining. 2004：168-177

[3] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C] // Proceedings of WWW-03, 12th International Conference on the World Wide Web. Budapest, HU, ACM, 2003：519-528

[4] 姚天昉，程希文，徐飞玉，等. 文本意见挖掘综述[J]. 中文信息学报，2008，22(3)

[5] Turney P D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [C] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002：417-424

[6] Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews [C] // Dale R, Paris C, eds. Proc. of the COLING/ACL 2006. Morristown：ACL，2006：483-490

[7] 任福继，等. Document for Ren-CECps 1.0 [OL]. http://al-www.is.tokushima-u.ac.jp/member/ren/Ren-CECps1.0/Ren-CECps1.0.html，2009

[8] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022

[9] Tsoumakas G, Katakis I. Multi-Label Classification：An Overview[J]. International Journal of Data Warehousing and Mining，2007，3(3)：1-13

[10] Tsoumakas G, Vlahavas I. Random K-Labelsets：an Ensemble Method for Multilabel Classification [C] // Proceedings of the 18th European Conference on Machine Learning (ECML2007). Warsaw, Poland，2007：406-417

[11] 孙艳，周学广，付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报：自然科学版，2013，1(49)

(上接第 16 页)

义运行剖面也没有简单的或可重复的方法。现在要具体给出样本空间$(\Omega,\mathcal{F},P)$及其上的随机变量ξ分布的构造理论，还要构造$\mathcal{F}$的子σ-代数$\mathfrak{B}$，从而研究$E(\xi|\mathfrak{B})$的条件分布，是有点困难的。但是，如果认真总结和思考数十年软件测试的经验(比如还可参看文献[9-15])，要在上面框架里做出一点成绩还是有希望的，例如我们就是在文献[17]的基础上，利用随机理论提出了随机软件错误定位方法，该方法的优势通过一些实例已经得到了验证，这也说明了我们提出的理论框架是有应用价值的。这正是我们今后研究的动力。

我们知道，在自然工程领域，只要有应用数学方法的都尽量运用，以保证工程质量。唯独软件工程，它属于社会-技术系统，再加上有时时间限制比质量限制更关乎在软件市场上的竞争力，所以软件工程界普遍认为使用数学论证方法不合算，即使 Java 语言提供了一些验证语句手段，它们在运行时往往也是"屏蔽"的(不过它们也确实起到了一定的验证作用)。因此我们强调开展软件测试关于审查和形式证明的研究，使其定位在形式系统不确定性和程序语言弱点上，这不仅方向明确而且更有实际可行性。

另外，概率论也很难"完整"地运用到软件测试中。即使统计测试也大都是关于软件可靠性方面的测试，它是基于统计学里的可靠性理论基础的。究其原因，软件测试里基于样本空间上的理论描述与基于经典概率论样本空间上的描述多少在细节上有微妙差别。为了突出差别，我们提出了样本空间的超滤模型和保测映射两种有别于经典的理论模型，希望它们更能满足软件测试随机理论的需要。对于完全的经典模型，我们也强调了条件期望理论，目前在证明等价划分的合理性上至少也看出该理论的作用。

参 考 文 献

[1] Schach S R. 软件工程—面向对象和传统的方法[M]. 邓迎春，韩松，徐天顺，等译. 北京：机械工业出版社，2007

[2] 莫绍揆. 数学基础[M]. 北京：高等教育出版社，1991

[3] Patton R. 软件测试[M]. 张小松，王钰，曹跃，等译. 北京：机械工业出版社，2007

[4] Pierce B C. 类型和程序设计语言[M]. 马世龙，眭跃飞，等译. 北京：电子工业出版社，2005

[5] Mitchell J C. 程序设计语言理论基础[M]. 许满武，徐建，袁宜，等译. 北京：电子工业出版社，2006

[6] Fenton N E, Pfleeger S L. 软件度量[M]. 杨海燕，赵巍，张力，等译. 北京：机械工业出版社，2004

[7] Winskel G. 程序设计语言的形式语义[M]. 宋国新，邵志清，等译. 北京：机械工业出版社，中信出版社，2007

[8] 王梓坤. 随机过程论[M]. 北京：科学出版社，1978：439-440，450

[9] Desikan S, Ramesh G. 软件测试-原理与实践[M]. 韩柯，李娜，等译. 北京：机械工业出版社，2009

[10] Andersson C, Runeson P. A Replicated Quantitative Analysis of Fault Distributions in Complex Software System [J]. IEEE Transactions on Software Engineering，2007，5(33)：273-286

[11] Cordy M, Classen A, Perrouin G, et al. Simulation-based abstractions for software product-line model checking [C] // Proceeding of the 2012 International Conference on Software Engineering. Zurich, Switzerland，2012：672-682

[12] 周毓明，徐宝文. 基于依赖结构分析的类重要性度量方法 [J]. 东南大学学报：自然科学版，2008，3(38)：380-384

[13] 王蓁蓁. 朴素模糊描述逻辑知识库构造及其朴素推理[J]. 应用科技，2012，39(6)：18-29

[14] Santelices R, Jones J A, Yu Yan-bing, et al. Lightweight Fault-Localization Using Multiple Coverage Types [C] // Proceedings of the 2009 IEEE 31st International Conference on Software Engineering. 2009：56-66

[15] Weimer W, Nguyen T, Goues C L, et al. Automatically Finding Patches Using Genetic Programming [C] // Proceedings of the 2009 IEEE 31st International Conference on Software Engineering. 2009：364-374

[16] Sommerville J. 软件工程[M]. 程成，陈霞，译. 北京：机械工业出版社，2008

[17] 王蓁蓁，徐宝文，周毓明，等. 一种随机 TBFL 方法[J]. 计算机科学，2013，40(1)：5-14