第 41 卷 第 4 期 计 算 机 工 程 2015 年 4 月
Vol. 41 No. 4 Computer Engineering April 2015

·人工智能及识别技术· 文章编号：1000-3428(2015)04-0171-05 文献标识码：A 中图分类号：TP18

基于图模型和多分类器的微博情感倾向性分析

黄 挺，姬东鸿

（武汉大学计算机学院，武汉 430072）

摘 要：为研究情感词对情感倾向分析的作用，提高微博情感分析性能，提出一种情感词图模型的方法，利用 PageRank 算法计算出情感词的褒贬权值，将其作为条件随机场模型的特征，预测具体语言环境下的情感词倾向。结合具体语境下的情感词倾向，利用支持向量机模型进行微博语料的主客观分类和情感倾向分类。实验结果表明，图模型构造的情感词典增加了具体语境下情感词倾向预测的准确性，具体语境下的情感词倾向预测对主客观分类和情感倾向分类有明显的改善。

关键词：图模型；情感词；条件随机场；支持向量机；网页排序算法；倾向性分析

中文引用格式：黄 挺，姬东鸿. 基于图模型和多分类器的微博情感倾向性分析[J]. 计算机工程，2015，41(4)：171-175.

英文引用格式：Huang Ting，Ji Donghong. Emotional Orientation Analysis of Microblog Based on Graph Model and Multiple Classifiers[J]. Computer Engineering，2015，41(4)：171-175.

Emotional Orientation Analysis of Microblog Based on Graph Model and Multiple Classifiers

HUANG Ting，JI Donghong

（Computer School，Wuhan University，Wuhan 430072，China）

【Abstract】 For the further research of the function of emotional words on emotional analysis and the improvement of microblog emotional analysis method，this paper proposes a research approach to construct emotional words graph model using relations between emotional words. The emotional value of appraisal calculated by PageRank algorithm and trained as the feature of conditional random field model so as to forecast the tendency of emotional words in specific situations，through which subjectivity classification and emotional tendency analysis of microblog can be made when integrated with Support Vector Machine（SVM） model. Experimental results show that emotional lexicon constructed by graph model enhances accuracy of the prediction of emotional word in specific situations which is also helpful for subjectivity classification and emotional tendency analysis of mircroblog.

【Key words】 graph model；emotional words；condition random field；Support Vector Machine（SVM）；PageRank algorithm；orientation analysis

DOI：10.3969/j.issn.1000-3428.2015.04.032

1 概述

随着网络的普及和热点舆论事件的网络化，人们更加倾向于在互联网上发表自己的个人观点和看法，使得互联网信息变得越来越丰富。这些信息不仅能帮助市场经济的参与者加强服务质量，而且还能为决策者提供更多的决策信息。

文本的情感倾向性分析是指对包含用户表示的观点、喜好、厌恶和情感等主观性文本进行情感分析。微博作为一种开放和即时性的信息传播媒介，不仅需要对其进行安全监管，防止可能对社会造成负面的影响，而且还需要对微博内容进行二次开发，挖掘其中的海量信息便于更好服务大众。最近这类的研究也比较多[1-2]，例如文献[3]重点考虑了连词对句子情感极性分析的影响，结合短语和连词分析句子情感极性。但系统依赖人工构建情感词典，并且需要人工构建连词规则，不具有领域适应能力。文献[4]基于情感词典扩展技术的网络情感倾向分

基金项目：国家自然科学基金资助重点项目(61133012)；国家自然科学基金资助面上项目(61173062)。

作者简介：黄 挺(1990－)，男，硕士研究生，主研方向：机器学习，数据挖掘；姬东鸿，教授、博士。

收稿日期：2014-05-23 **修回日期**：2014-06-26 **E-mail**：timhuangwhu@gmail.com

析,应用 HowNet[5] 和 NTUSD[6] 2 种资源对现有情感词典进行扩展,建立了一个新的、具有倾向程度的情感词典,但是建立的情感词典不能完全反应特殊的微博语言环境。文献[7]也提出一种融合最大熵模型和支持向量机模型两者预测结果用于识别主观句和褒贬极性分类问题的方法。

针对上述问题,本文提出基于图模型识别情感词语的情感倾向[8]和计算情感倾向度的方法,在此的基础上利用条件随机场模型[9]训练和预测具体语境下的情感词倾向[10],并结合支持向量机模型预测主客观句和褒贬极性问题[11]。

2 情感词典的构建和优化

2.1 图模型

图模型即表示为点和边的集合,点代表着情感词,边代表情感词对的潜在情感倾向相似关系,并用出现频率表示其权值。

为了寻找潜在相似情感倾向的词语对,需要在分词和依存分析的基础上进行如下步骤:

(1)通过依存关系分析,对于语法结构中依存于同一对象的修饰词,认定为潜在相似情感词对。例如“坑爹的国足,实在没劲!”,其依存分析如图 1 所示,虽然“坑爹”和“国足”之间没有直接的依存关系,但是“的”没有实际的语义作用,相当于“坑爹”和“国足”存在间接的依存关系,“坑爹”和“没劲”都修饰“国足”且都有相同的依存关系,该语境下这对词有相同的情感倾向。



图 1 依存关系示例

需要注意的是转折连词和转折副词可能会改变词语对的相似性,此处采用的方法是出现就直接忽略,不作统计。

(2)对于同一个句子中多个并列结构短语的修饰词,也认为是潜在相似情感词对。例如“政客们的虚伪,统治者们的残忍,民众的盲从…”,在该语境下可认为虚伪、残忍和盲从有着相似的情感倾向。

(3)同一句子中通过共现关系抽出用并列连词连接的潜在相似情感词对。例如“今年结了婚,又买了房子,感觉特别幸福和快乐!”,可以大致推断“幸福”和“快乐”这组词语在特定的环境下有着相同的情感倾向。

最终通过统计潜在相似情感词对的出现频率,如果超过 X 次则加入到图模型中,并用出现频率表示边的权重。

通过新浪微博的 API 得到了大约 10 万条微博数据进行上述规则的抽取,最终形成了如图 2 所示的图模型。

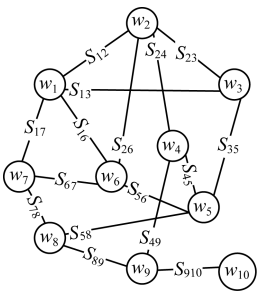


图 2 图模型部分示例

2.2 种子词

对上述微博语料进行处理之后(分词,去掉停用词,剪枝等),对潜在相似情感词进行 TF-IDF 统计,具体公式如下:

$$TF-IDF(W)=\sum_{x=0}^{N}TF_x(W)\times IDF_x(W) \quad (1)$$

为了提高统计效果和精度,将同一话题的微博评论形成一个微博集合,其中,N 表示微博集的数量,词频(Term Frequency,TF)表示词语 w 在第 N 个微博集合中的出现频率和语料中总出现频率的比值,逆向文档频率(Inverse Document Frequency,IDF)表示微博集合数 N 和出现词语 w 的微博集合数量比值的对数。

在 TF-IDF 值较高的词语中人工挑选了 40 个情感词作为情感基准词[11],即非常典型而且确定只含褒义或者贬义的情感倾向词。褒贬义情感基准词如下:(1)褒义基准情感词:大气,出色,完美,漂亮,幸福,先进,优秀,文明,美丽,真实,可爱,健康,认真,美好,和平,赞美,积极,欢乐,顽强,开朗。(2)贬义基准情感词:腐败,非法,恶意,诽谤,浪费,变态,漏洞,欺诈,野蛮,陷阱,堕落,自大,贪污,色情,虚假,压榨,恶魔,谎言,变态,造假。

2.3 情感倾向权值的计算

利用 PageRank 算法对形成的情感词网络进行情感词评分。主要如下:

(1)把每个待预测情感词赋予正负面 2 个情感度权值,分别表示它们正负面的情感强烈度,负值代表贬义倾向,正值代表褒义倾向。绝对值越大表示情感倾向越强,正负面情感词权值在迭代过程中相互独立。即:

$$Weight(w)=(Postive(W),Negetive(W)) \quad (2)$$

(2)情感基准词只会向周围传递它唯一的褒义或者贬义情感倾向权值,另一倾向权值始终设为 0。

例如“快乐”只有正面的含义,在负面倾向权值计算过程中,其负面权值始终为 0,相当于在负面的计算过程中忽略了“快乐”这个词对周围相似情感词的影响。而正面倾向权值计算过程方法不变。

$$\boldsymbol{A}=\begin{pmatrix}F_{11} & F_{21} & \cdots & F_{n1}\\ F_{12} & F_{22} & \cdots & F_{n2}\\ \vdots & \vdots & & \vdots\\ F_{1n} & F_{2n} & \cdots & F_{nn}\end{pmatrix} \tag{3}$$

$$P_i=\boldsymbol{A}P_{i-1} \tag{4}$$

其中:

$$P=(Postive(W_1),Postive(W_2),\cdots,Postive(W_n))$$

$$N_j=\boldsymbol{A}N_{j-1} \tag{5}$$

其中:

$$N=(Negetive(W_1),Negetive(W_2),\cdots,Negetive(W_n))$$

其中,式(3)表示待预测情感词之间关系出现的统计频率;式(4)表示待预测情感词正面的权值向量 P 第 i 次迭代的过程;式(5)表示待预测情感词负面的权值向量 N 第 j 次迭代的过程。

初始化:

$$P_0=(1/M,1/M,\cdots,1/M)$$

$$N_0=(-1/M,-1/M,\cdots,-1/M) \tag{6}$$

其中,M 表示待计算的情感词数量,以上对情感基准词某一倾向归零的处理相当于在该倾向计算过程中忽略该词在图中的影响作用,不会影响图模型的收敛速度,所以通过多次的迭代之后,该方法最终会得到稳定的正负权值向量。

2.4　情感词典优化

在之前计算结果的基础之上对情感词的情感权值进行标准化,通过式(7)、式(8)将其转换至[-1,1]区间。

$$Postive(W_i)=\frac{2}{1+e^{\frac{1}{P[i]}}} \tag{7}$$

$$Negetive(W_i)=-\frac{2}{1+e^{-\frac{1}{N[i]}}} \tag{8}$$

其中,$P[i]$ 和 $N[i]$ 分别表示情感词的 W_i 正负面情感权值,且不为 0。式(7)、式(8)2 个挤压函数将初始的权值调整至[-1,1]区间,由于挤压函数为 S 型增长函数,对少量初始权值绝对值较大或者较小的情况都有较好的控制作用。S 型函数中间坡度较大,对两者之间的大量中间情况有较好的区分度,所以此处的转换具有较大的意义。

最后对情感权值的绝对值低于 $\theta(0<\theta<1)$ 的情况,将情感权值归 0,视为无正面或负面的情感倾向,当其正负情感权值均为 0 时,将该词从情感字典删除,以提高图模型的效果。

3　微博情感倾向分析

3.1　特征选择

由于微博语句简短,口语化较多,随意性比较强,以及结构的不严谨性等特点,因此在特征提取时不仅要考虑到普通文本特征,还要提取一些专门针对微博语境的特征,具体的特征如表 1 所示。

表 1　特征抽取项

序号	特征描述	示例和解释
1	微博表情	[怒][嘻嘻]等微博表情
2	词性	每个词对应的词性
3	情感词的情感倾向和情感权值	通过分类器得出的情感词在具体语境下的情感倾向
4	特殊句式	疑问、反问、否定、双重否定等特殊句式
5	暗含情感的标点符号	“…”、“!!!”和“???”等等
6	网络常见口语话用词	“给力”等就是网络较为常见词语
7	情感词附近表示程度的修饰词	“有点”、“太”、“不一定”、“完全”和“之极”等
8	前后句的承接关系词	“但是”、而且、“却”和“可是”等

3.2　褒贬极性分类

褒贬极性分类的主要流程如下:

(1)在前文建立的情感词典基础上,利用条件随机场模型(Condition Random Field,CRF)对具体语境下的情感词进行情感倾向预测。

具体语境下的情感词情感倾向会随着语境的变化而出现正负面变化,所以情感词在语境中的实际倾向也可以作为褒贬分类的特征之一,并且在实验中也证实了其重要性,对结果有较明显的改善和提高。处理流程如图 3 所示。

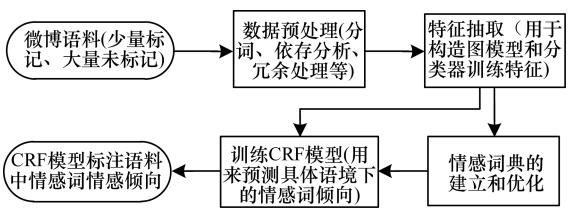


图 3　CRF 情感词标注处理流程

具体的 CRF 训练预测流程如下:

1)根据图模型和 PageRank 算法构建情感词典

和计算情感倾向权值。

2）预测过程：

①一条微博训练语料；

②抽取该微博特殊句式、特殊符号、情感词和特殊微博表情等特征；

③提取情感词典中出现在该微博里的情感词信息；

④将之前信息根据特征模板组成特征序列，特征序列包括表 2 中微博符号、表情和句式等在句子中的相对位置和特征编号，情感词典中出现在该微博里的情感词的情感倾向权值和在微博中相对位置，标注序列为人工标注的情感词情感倾向。

3）将特征序列和标注序列整理成语料并作为输入序列训练 CRF 模型。

（2）结合语言特征和情感词具体倾向，由支持向量机模型（Support Vector Machine，SVM）进行主客观分类和情感倾向分类。

把 CRF 对情感词的倾向预测结果也作为特征，结合特征抽取的特征集，对支持向量机模型进行训练分类，当分类点到训练模型的超平面距离为正值时，即为褒义的情感倾向微博，反之为贬义的情感倾向微博。

3.3 主客观分类

有时可以直接对文本进行情感倾向性分类，但是如果能预测文本的主客观情况，就可以在分类之前排除客观语料，一定程度上提高情感倾向分析的性能，本文在主客观文本分类上还是使用了之前的模型和方法，只是为了平衡数据结构，在测试集中添加了部分手工标注的客观微博数据。经过主客观分类之后的数据仅包含主观的情感倾向数据，只需要进行二元分类就能得到最终的结果。具体流程和前面类似，不再赘述。

4 实验及结果分析

4.1 语料预处理

在使用微博语料之前要对其进行语料预处理。本文的处理主要包括以下 7 点：（1）“#话题#”过滤；（2）“@昵称”过滤；（3）链接过滤；（4）分词；（5）停用词处理；（6）网络新词、热词处理；（7）对语料进行依存分析。

4.2 训练语料和测试数据

本文通过新浪微博 API 获取大约 10 万条未标记语料作为训练图模型的数据，再加上计算机学会中文情感倾向评测的标记数据选择 6 740 条作为训练语料，其中，正面 3 532 条；负面 3 208 条。另外还选择 5 021 条作为测试语料，其中正面 2 323 条，负面 2 198 条，另外还人工标注了 2 000 条客观微博数据作为主客观分析的语料。

4.3 结果分析

实验结果与分析如下：

（1）情感词的正确率和覆盖率，如图 4 所示。

在情感权值计算的过程中会对情感绝对权值小于阈值 θ 的情感词舍去，降低图模型在计算迭代的过程中产生的干扰。图 4 显示了不同阈值对情感词典正确率和覆盖率的影响。

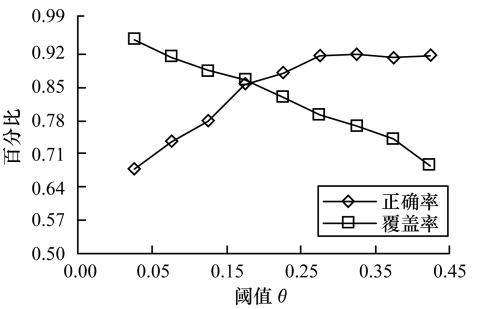


图 4 阈值 θ 的影响

覆盖率是指测试文本中由图模型得出的情感词出现频率和情感词出现的总频率比值。从图 4 可以看出：阈值越大正确率越高，最后正确率会逐渐稳定到一个定值。而覆盖率会随着阈值上升而逐渐降低，所以综合考虑将阈值设置在 0.18 比较合适。

（2）潜在相似情感词词频阈值 X 对最终倾向分析准确性的影响，如图 5 所示。

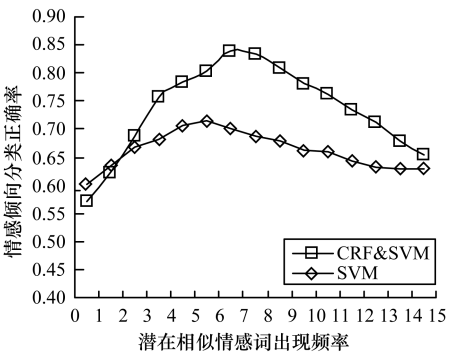


图 5 阈值 X 对结果的影响

潜在相似情感词词频的阈值会影响到图模型中节点和边的数量，从而影响最终的情感词典和权值，进而影响结果，从实验数据来看，最后的正确率首先会受到词频的增大而增加，到达峰值后，由于词频过大导致限制严格，最终效果反而下降。

（3）微博特征对情感倾向分类的影响，如图 6 所

示。图 6 表示对应表 2 各特征项当其缺省时的最终正确率。可知情感权值、特殊句式和网络流行语等对效果有更显著的影响作用。

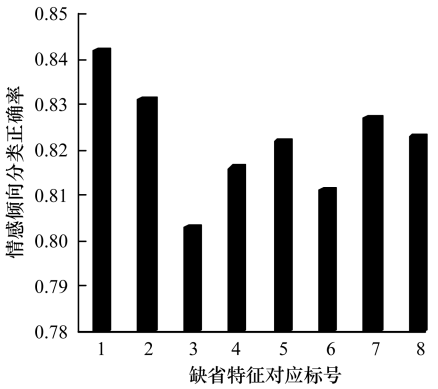


图 6　各缺省特征对结果的影响

(4)主客观分类结果如表 2 所示。

表 2　主客观分类结果

结果类型	正确率	召回率	F 值
SVM	0.740	0.681	0.709
CRF&SVM(1)	0.753	0.657	0.701
CRF&SVM(2)	0.808	0.816	0.812
平均值	0.727	0.615	0.647
最高值	0.835	0.959	0.892

其中,CRF&SVM(1)表示使用台湾大学情感词典的实验结果;CRF&SVM(2)表示使用本文方法建设情感词典得出的结果,平均值和最高值均参考 2013 年 CCF 中文情感评测的结果。

(5)情感倾向分析的结果如表 3 所示。

表 3　情感倾向分类结果

结果类型	正确率	召回率	F 值
SVM	0.725	0.651	0.686
CRF&SVM(1)	0.752	0.658	0.702
CRF&SVM(2)	0.852	0.754	0.801
平均值	0.704	0.460	0.546
最高值	0.902	0.850	0.850

从表 3 数据可以看出:CRF 预测情感词倾向加上 SVM 情感倾向分类的方法比仅仅使用 SVM 模型效果明显有提升,这是因为情感词典只是整个微博环境下情感词的倾向分布,而 CRF 模型能够利用特征之间的序列关系和微博情感词的倾向分布较好的预测具体语境下的情感词倾向,这对之后的情感倾向分类有比较积极的作用,一定程度上提高了倾向分类预测的准确率。

另外本文方法通过使用普通情感词典和本文建立的情感词典的比较可知利用微博数据建立的情感词典更具有领域适应性,分类预测效果更好。但是实验效果离最高值还有一定的距离,还需要不断努力和改进。

5　结束语

本文利用情感词之间的相互关系,建立一个包含情感词及其关系的图模型,采用 PageRank 算法计算情感词的正负情感倾向权值,并结合其他特征利用 CRF 模型进行具体语言环境下的情感词倾向预测,并以此为基础采用 SVM 模型来预测句子的主客观性和褒贬极性。在普通的微博语料中,本文方法都获得较好的效果。下一步将在本文情感词典建设的基础上丰富词典和提高精确度,继续对使用的分类方法进行优化,并尝试结合其他算法进行情感倾向分析。

参考文献

[1] 周立柱,贺宇凯,王建勇. 情感分析研究综述[J]. 计算机应用,2008,28(11):2725-2728.
[2] 周胜臣,瞿文婷,石英子,等. 中文微博情感分析研究综述[J]. 计算机应用与软件,2013,30(3):161-164.
[3] Meena A,Prabhakar T,Amati G,et al. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis[C]//Proceedings of Advances in Information Retrieval. Berlin,Germany:Springer,2007:573-580.
[4] 杨　超,冯　时,王大玲,等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统,2010,4(4):691-695.
[5] 朱嫣岚,闵　锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报. 2006,20(1):245-251.
[6] 杨昱昺,吴贤伟. 改进的基于知网词汇语义褒贬倾向性计算[J]. 计算机工程与应用,2009,45(21):91-93.
[7] 刘志广,董喜双,关　毅. 基于分类器融合的中文微博情感倾向性研究[C]//2012 年第一届自然语言处理与中文计算会议论文集. 北京:[出版者不详],2012.
[8] 傅向华,刘　国,郭岩岩,等,中文博客多方面话题情感分析研究[J]. 中文信息学报,2013,27(1):47-55.
[9] 杨　源,林鸿飞. 基于产品属性的条件句倾向性分析[J]. 中文信息学报,2011,25(3):86-92.
[10] 王　勇,吕学强,姬连春. 基于极性词典的中文微博客情感分类[J]. 计算机应用与软件,2014,31(1):34-37.
[11] 赵　煜,蔡皖东,樊　娜,等. 利用词汇分布相似度的中文词汇语义倾向性计算[J]. 西安交通大学学报,2009,20(06):33-37.
[12] 闻　彬,何婷婷,罗　乐,等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学,2010,37(6):261-264.

编辑　索书志