基于短语模式的文本情感分类研究*)

李 钝[1,2] 曹付元[3] 曹元大[1] 万月亮[1]
(北京理工大学计算机科学技术学院 北京 100081)[1] (山西大学现代教育技术中心 太原 030006)[2]
(山西大学计算机科学技术学院 太原 030006)[3]

摘 要 文本倾向识别的研究在诸多领域有着广阔的发展前景,短语模式的文本情感分类是问答系统、信息安全、网上调查等研究的基础。本文从语言学角度出发,首先,分析词典中对词语义定义的特点,采用"情感倾向定义"权重优先的计算方法获得短语中各词的语义倾向度,然后分析短语中各词组合方式的特点,提出中心词概念来对各词的倾向性进行计算来识别短语的倾向性和倾向强度。实验表明,本文的方法对短语的倾向分类识别效果较好,可为更大粒度的文本倾向识别打好基础,具有一定的实用价值。

关键词 文本分类,情感倾向,语义倾向度,知网,短语结构,中心词

Text Sentiment Classification Based on Phrase Patterns

LI Dun[1,2] CAO Fu-Yuan[3] CAO Yuan-Da[1] WAN Yue-Liang[1]
(School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081)[1]
(Modern Education Technology Center, Shanxi University, Taiyuan 030006)[2]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006)[3]

Abstract The research of the text sentiment classification has the broad prospect and the related classification on phrases is the basic in QA system, information security and online investigation. Based on the linguistics, the paper analyzes the characteristics of the semantic definition in HowNet, calculates the orientation similarity of words in the phrase based on the sentiment weight priority and puts forward the concept of center word to calculate the orientation of the phrase according to the combination of the words in the phrase. The experiments show that it has better result in orientation recognition and some practice value that it can ground text orientation recognition in the research on larger granularities.

Keywords Text classification, Sentiment orientation, Semantic orientation, HowNet, Phrase structure, Center word

1 引言

随着互联网的安全发展和人们对信息需求的日益增强,面对庞大而且急剧膨胀的文本信息,要准确地找到自己所需要的信息越来越困难,通过搜索引擎等工具可以找到一些相关的信息,但是要找到跟自己观点相同的几乎是不可能的。文本情感分类作为一种新兴技术逐渐受到人们关注和重视,它能够对文档进行自动分析获取与人们兴趣观点相同的信息,将会对信息检索、信息过滤、反动信息的拦截和网上调查等研究提供便利。

2 相关研究

自从上世纪90年代以来,文本情感倾向性的研究在国内外受到了普遍的关注,并迅速发展起来。研究者从词、短语到篇章的研究如火如荼。例如,Wiebe[1]从一些分散的形容词和动词中学习线索,对动词进行N-grams分析,识别句子的褒贬性,从而对社论进行分类,Riloff[2]通过两种不同的Bootstrapping算法和一组种子词提取1000多个情感名词建立情感分类器,然后采用情感名词、谈话特征和情感线索进行贝叶斯分类;Turney[3]提出了利用统计信息对单词进行语义倾向判断的新方法,他所处理的对象可以是形容词、副词、名词、动词,其多角度问答系统(MPQA)采用观点标注语料库,其中角色的情感根据其个体状态、谈话事件、表达的主观因素来标识,要求判断上下文中所有情感的表达,所采用的PREV线索由人工建立且不局限于固定的词表或特定词性;Wilson[4]采用自定义的44个特征进行模式匹配获取多层嵌套的从句的倾向性,识别主客观的级别。

大部分的倾向识别的工作主要集中在词汇级或文档级,但是诸如问答系统、自动摘要、挖掘产品评论、信息安全等都需要的短语级或者句子级的情感倾向分析却研究甚少。Hatzivassiloglou和McKeown[5]从一个大的未标示的文档集中提取并分析形容词对(由and,or,but,either-or,或neither-nor连接)生成词汇间的同义或反义倾向的连接图,用聚类的方法将词聚成褒义和贬义两类,来预测具有主观性的形容词的倾向,该方法可行的根本原因是形容词的连接是受这些形容词倾向在语言学上的限制所影响的,例如,and通常连接相同倾向性的形容词,而but连接倾向性相反的形容词;Casey[6]从文本中提取出诸如"very good"和"not terribly funny"之类的形容词短语的特征评价组(Appraisal Groups)并进行分析获取其倾向性;Takamura[7]建立潜在变量模型(Latent Variable Models)对"名词+形容词"模式的短语进行倾向分类。由于中英文表述语义的方式习惯的不同,而且这些方法

*)基金项目:太原市科技局项目。李 钝 讲师,博士研究生,研究方向为信息过滤、内容安全;曹付元 讲师,博士研究生,研究方向为数据挖掘、机器学习;曹元大 教授,博士生导师,研究方向为计算机网络、信息安全;万月亮 博士研究生,研究方向是信息过滤。

仅针对短语中的某些个别类型进行研究，不能很好地为更大粒度的倾向分类识别打好基础。本文从语言学角度出发，首先，分析词典中对词汇语义定义的特点，采用“情感倾向定义”权重优先的计算方法获得短语中各词的语义倾向度，然后分析短语中各词组合方式的特点，提出中心词概念计算短语的倾向性和倾向强度，对文本进行情感分类。

3 基于语义学的短语倾向计算

作为文本语义倾向分类的研究基础，短语是词和词按照一定方式组合起来的语言单位，它的意义不仅包括内部各词的语义，还有各词之间的组合方式，因此本文根据短语内部各词的语义和词之间的结构信息确定短语的倾向性和倾向强度。

3.1 词的情感倾向度

3.1.1 HowNet 中义项之间的倾向相似度

在 HowNet 中，词有多个义项表述不同概念，而每个义项 S_k 又是由多个义原或具体词 $p_{k1},\cdots,p_{kt}$ 描述其特征，不同义项 S_1 和 S_2 之间的语义倾向相似度定义如下：

$$sim(S_1,S_2)=\frac{1}{t_1 t_2}\sum_{i=1}^{t_1}\sum_{j=1}^{t_2}\frac{1}{2^{t_1-i+1}}sim(p_{1i},p_{2j}) \quad (1)$$

其中，t_1 和 t_2 分别是 S_1 和 S_2 的义原个数，$\frac{1}{2^{t_1-i+1}}$ 是 S_1 中义原 p_{1i} 的位置权重，义原的相似度 $sim(p_{1i},p_{2j})=\frac{\alpha}{d+\alpha}$，$d$ 是 p_{1i} 和 p_{2j} 的路径距离，α 是可调节的参数。

3.1.2 词的情感倾向度

为了计算出普通词的情感倾向度，从 HowNet 中选取出一组强烈褒贬倾向且具有代表性的词组成种子集，根据词 w 与种子集中每个词的语义倾向相似程度来确定词 w 的语义倾向值。定义种子集为 seedset$=\{PP, PN\}$，其中，PP 指褒义种子词集，PN 为贬义种子词集。词 w 与褒义种子集中的各个词联系越紧密，则该词的褒义倾向越强烈；与贬义种子集中的各个词联系越紧密，则该词贬义倾向越明显。

词 w 的语义倾向值定义如下：

$$o(w)=\frac{1}{K}\sum_{k=1}^{K}sim(w,pp_k)-\frac{1}{L}\sum_{l=1}^{L}sim(w,pn_l) \quad (2)$$

其中，$pp_k\in PP$，$pn_l\in PN$，K 和 L 分别为褒义种子集和贬义种子集中种子词的个数。设置阈值 $\theta(\theta\geqslant 0)$，$o(w)>\theta$ 表明该词是褒义的，$o(w)<\theta$ 表示该词是贬义的，$|o(w)|\leqslant\theta$ 则表示该词是中性词，$o(w)$ 数值大小则代表词 w 褒贬倾向强度。

3.2 使用互信息计算词之间的关联强度

互信息是用于表征两个变量之间的相关性的一个度量方式。对于词 w_1 和 w_2，其互信息记为 $MI(w_1,w_2)$，计算公式如下：

$$MI(w_1,w_2)=\log_2\left(\frac{p(w_1\&w_2)}{p(w_1)\cdot p(w_2)}\right) \quad (3)$$

其中，$p(w_1\&w_2)$ 是词 w_1 和 w_2 的有序同现的概率。

3.3 短语的倾向识别

3.3.1 短语中的语义结构

语言中的各个词之间是相互联系和相互制约的关系，短语中各个词之间可以是实词和实词的组合，也可以是实词和虚词的组合，相邻词之间可以是并列存在相互独立的关系，也可以互为修饰和被修饰关系。具有一定实在含义并且能表现语义倾向性的短语一定是以名词、动词、形容词、副词等实词为中心词出现的，常见的短语类型有偏正、主谓、动宾、并列、连动等，根据短语内部中心词的个数可以分为两类：

1)一个中心词(修饰词和功能词在前或在后，增强或减弱中心词的倾向性)：偏正、主谓、动宾、介宾、补充、的字、所字

举例：偶像 的 亲和力，聪明 女孩，很 好

2)多个中心词(内部各个词对于短语语义的贡献度是相同的)：并列、连动、同位、兼语

举例：透露 出 贪婪 和 邪恶 的 眼神，聪明 伶俐 的 姑娘

有些短语内部词之间有明显的并列或转折连接词，如“和”、“与”、“又”等，而有些直接组合，有些用顿号或逗号隔开，需要从语义和词性等方面进行分析才能够判断短语内部的结构组成，如表 1 所示。

表 1 短语内部语法结构

	特征	例子
一个中心词	形＋名	聪明女孩
	名＋动，名＋形	小王喜欢
	动＋名，动＋形	爱干净
	名＋“的”＋名	偶像的亲和力
	动/形＋“得”＋动/形	过得开心
	程度词＋形/副，形/副＋程度词	很好
	否定词＋动/名/形	不喜欢
多个中心词	并列或转折连词	贪婪和邪恶
	形＋形，名＋名，动＋动	聪明伶俐

3.3.2 短语的语义倾向值计算

如 3.3.1 所介绍，按中心词的个数对短语分类，不同类别采取不同的倾向值计算策略，该倾向值受短语中各词的语义倾向度和相互之间的语义关联程度所影响。以两个词组成的短语为例，令短语 $phrs=\{w_1w_2\}$，其中 w_1 和 w_2 为两个汉语词。

1) 一个中心词的计算方法如表 2 所示。

表 2 一个中心词的短语倾向计算

w_1	w_2	$o(phrs)$	说明
$o(w_1)\neq 0$	$o(w_2)\neq 0$	$o(w_1)o(w_2)MI(w_1,w_2)$	两词均有情感倾向
$o(w_1)=0$	$o(w_2)\neq 0$	$o(w_2)MI(w_1,w_2)$	有一词为中性词
$o(w_1)\neq 0$	$o(w_2)=0$	$o(w_1)MI(w_1,w_2)$	
$o(w_1)\neq 0$	$w_2\in NEG$	$-o(w_1)$	有一词为否定词
$w_1\in NEG$	$o(w_2)\neq 0$	$-o(w_2)$	
$o(w_1)\neq 0$	$w_2\in DGR$	$g(w_2)\ o(w_1)$	有一词为程度词
$w_1\in DGR$	$o(w_2)\neq 0$	$g(w_1)\ o(w_2)$	
$o(w_1)=0$	$o(w_2)=0$	0	两词均为中性词

2)多个中心词的计算方法

$$o(phrs)=(o(w_1)+o(w_2))\cdot MI(w_1,w_2) \quad (4)$$

4 实验及分析

实验采用了两个语料库 D1 和 D2，语料库 D1 包括 324 篇关于超女的评论，同一篇评论中可能包括对多个人的褒贬不一的评价，但由于作者对不同超女的爱憎较分明常采用一些感情色彩强烈的短语词汇，故适合本文针对短语进行情感识别的实验要求；语料库 D2 来自于两全其美论坛中的“焦点动态”和“军事”等热门新闻话题，包括 50 个热门话题，共 9000 多个帖子，篇幅长短不一，但每个帖子对于话题的观点都很明确，因此用于进行基于短语的文档情感倾向分类实验。

4.1 短语倾向识别

文本倾向识别，比较常见的方法有传统的分类方法和采用评价组 AG(Appraisal Groups)进行情感分析的方法[6]。本

文实验把这两种方法与本方法作比较，从 D1 语料库中识别出有情感倾向的短语，其中，识别出的短语包括褒义和贬义短语，中性短语属于未识别出情感倾向的短语不包括在内。传统的文本分类方法需要事先标注好一些短语作为学习样本，然后使用 SVM 方法构造一个两类分类器对测试集中的短语进行分类。AG 方法可以很好地描述诸如“很幸福”和“不是很幸福”之类短语的情感识别，但是它只能识别短语内部各词之间的“修饰/被修饰”关系，对其他的处理能力较弱。本文在其基础上从词法分析的角度分析短语内部结构，取得较好的效果。从表 3 可以看出，基于 AG 的方法对多中心词的短语识别能力较差，传统分类方法介于两者之间，但是传统分类方法需要人工标注训练样本，针对不同主题构造不同的分类器，通用型较差，而本文的方法从短语的两要素出发，利用 How-Net 对短语中的各个词语义的准确定位，然后基于词法分析对其中各个词之间的关系进行分析，能够较准确地进行倾向分类。

表 3　三种方法对 D1 语料库中短语的分类实验比较

语料库 D1	传统方法	AG 方法	本文方法
一个中心词	88.3%	86.2%	93.5%
多个中心词	89.1%	38.6%	95.7%
平均值	88.7%	62.4%	94.6%

4.2　基于短语识别的文档倾向分类

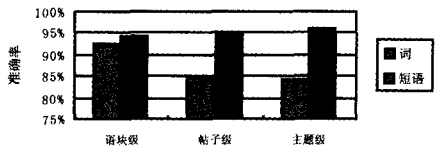

图 1　使用词和短语进行文档倾向分类比较

文档情感分类主要有两种方法：一种是基于文档中的各个词的同现频率进行褒贬倾向的文档分类；另一种是把所有词分成两类，然后统计文档中两类词的总数进行分类。为了更好地比较本文短语倾向识别方法的优劣性，采用第二种方法对语料库中的文档进行情感倾向分类。现 D2 语料库中包含 50 个主题，9000 多帖，有 947 万字，经过 ICTCLAS 分词获得 617 万多个词块，从中提取实词(包括 n，v，a，d，ad，vd，vn 等)，其中有情感倾向的词为 185,126 频次，短语 101,634 频次，采用本文的方法分别获取词和短语的倾向性，根据词和短语的褒贬性获得各帖的褒贬倾向性，然后针对同一主题统计褒贬帖数将主题进行分类，图 1 为使用词和短语对 D2 语料库中的主题进行情感倾向分类的准确率比较。

从图 1 可以看出，在语块级的倾向分类实验中，基于词的比基于短语的分类准确率略低，因为两个实验均基于语块本身的语义进行比较，没有考虑上下文环境对该语块的影响以及语块对文档主题的影响，所以各个语块的准确率都较高且差距不大；但在帖子级和主题级，这种状况就有所改变，基于词的和基于短语的分类准确率差距加大，主要原因是短语相对于词来说，语义粒度变大，对主题体现的能力更强，准确率更高。

结论　大部分的倾向识别的工作主要集中在词汇级或文档级，但诸如问答系统、摘要提取、挖掘产品评论等都需要句子级或者短语级的倾向分析，而这些方面的研究却较少。本文从语言学角度出发，首先分析词典中对词汇语义定义的特点，采用“情感倾向定义”权重优先的计算方法获得短语中各词的语义倾向度，然后分析短语中各词组合方式的特点，提出中心词概念来对各词的倾向性进行计算来识别短语的倾向性和倾向强度。实验表明，本文的方法对短语的倾向分类识别效果较好，可为更大粒度的文本倾向识别打好基础，具有一定的实用价值。但是文档倾向性还受文档结构等诸多因素影响，还需要通过中心句中心段的识别来提高文档级的分类精度，有待进一步更深入细致的研究。

参 考 文 献

1　Wiebe J，Wilson T，Bell M. Identifying collocations for recognizing opinions[A]. In：Proc. ACL-01 Workshop on Collocation：Computational Extraction，Analysis and Exploitation[C]，2001

2　Riloff E，Wiebe J，Wilson T. Learning Subjective Nouns using Extraction Pattern Bootstrapping[A]. In：Conf. on Natural Language Learning (CoNLL)[C]，2003. 25～32

3　Turney P，Littman M. Measuring praise and criticism：Inference of semantic orientation from association[J]. ACM Transactions on Information Systems，2003，21(4)：315～346

4　Wilson T，Wiebe J，Hwa R. Just how mad are you? Finding strong and weak opinion clauses[A]. In：Proceedings of 21st Conference of the American Association for Artificial Intelligence (AAAI-04)[C]. US，2004

5　Hatzivassiloglou V，McKeown K R. Predicting the semantic orientation of adjectives[A]. In：Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)[C]，1997. 174～181

6　Whitelaw C，Garg N，Argamon S. Using Appraisal Groups for Sentiment Analysis[A]. In：Proceedings of the 14th ACM international conference on Information and knowledge management [C]，Bermen，Germany，2005. 625～631

7　Takamura H，Inui T，Okumura M. Latent Variable Models for Semantic Orientations of Phrases[A]. In：Proc. of the 11th Conference ofthe European Chapter of the Association for Computational Linguistics (EACL-2006)[C]，2006

(上接第 108 页)

动态调整反馈信息与量化门限，既可节省系统的反馈开销，同时也能提高系统的频谱效率，所有的反馈与量化参数都可做离线计算，并以表的形式存储在用户端以供查询，降低了实时运算的复杂度，使该算法更具实用性。

参 考 文 献

1　Shen Zukang，Andrews J G. Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. IEEE Transactions on Wireless Communications，2005，4(6)：2726～2737

2　Cimini L J，Daneshrad B，Sollenberger N R. Clustered OFDM with transmitter diversity and coding. IEEE GLOBECOM，1996. 703～707

3　Svedman P，Wilson S K. A simplified opportunistic feedback and scheduling scheme for OFDM. IEEE VTC，2004. 1878～1882

4　Wong C Y，Cheng R S，Letaief K B，et al. Multiuser OFDM with adaptive subcarrier，bit and power allocation. IEEE J Select Areas Commun，1999，17(6)：1747～1758

5　Goldsmith A J，Chua S. Vairable rate variable power MQAM for fading channels. IEEE Transactions on Communications，1997，45(10)：1218～1230

6　Liu Qingwen，Zhou Shengli，Giannakis G B. Cross-Layer Scheduling With Prescribed QoS Guarantees in Adaptive Wireless Networks. IEEE journal on selected areas in communications，2005，23(5)：1056～1066