基于语义理解的文本情感分类方法研究

闻 彬[1,2,3] 何婷婷[1,2] 罗 乐[1,2] 宋 乐[1,2] 王 倩[1,2]
(华中师范大学计算机科学与技术系 武汉 430079)[1]
(国家语言资源监测与研究中心网络媒体分中心 武汉 430079)[2] (咸宁学院计算机学院 咸宁 437100)[3]

摘 要 文本情感分类方法在信息过滤、信息安全、信息推荐中都有广泛的应用。提出一种基于语义理解的文本情感分类方法,在情感词识别中引入了情感义原,通过赋予概念情感语义,重新定义概念的情感相似度,得到词语情感语义值。分析语义层副词的出现规律及其对文本倾向性判定的影响,实现了基于语义理解的文本情感分类。实验表明,该方法能有效地判定文本情感倾向性。

关键词 文本情感分类,情感倾向性,语义理解

Text Sentiment Classification Research Based on Semantic Comprehension

WEN Bin[1,2,3] HE Ting-ting[1,2] LUO Le[1,2] SONG Le[1,2] WANG Qian[1,2]
(Department of Computer Science, Huazhong Normal University, Wuhan 430079, China)[1]
(Monitor and Research Center for National Language Resource Network Multimedia Sub-branch Center, Wuhan 430079, China)[2]
(Department of Computer Science, Xianning College, Xianning 437100, China)[3]

Abstract Text sentiment classification is used widely, such as information filtering, information security and information recommendation. This paper proposed an improved method based on semantic comprehension. In this paper, sentiment primitive was introduced into sentiment terms identification, and then semantic values of phrases were obtained. Furthermore, this paper further analysed adverbs and its influence on identification of text orientation in the semantic level and achieved the text sentiment classification. The experimental results show that the proposed approach is suitable for judging sentiment orientation.

Keywords Text sentiment classification, Sentiment orientation, Semantic comprehension

随着万维网的飞速发展和信息资源的高速膨胀,网络作为知识获取、信息交流的主要工具已经逐渐发展成为一种开放式平台,人们可以自由地在各种论坛、博客上发表自己的意见和观点。文本倾向性分析可以用来处理各种断言或评论,分析其中包含的情感和态度等。目前,文本情感分类作为信息处理技术的一个研究重点被广泛应用在信息检索、信息过滤、信息安全、自动文摘等领域。

在自然语言处理研究中,对于文本情感分类的研究,通常是先针对词语进行情感识别。早在1997年,Hatzivassiloglou和McKeown[1]就尝试使用监督学习的方法对英文的词语进行情感语义倾向性判别;Turney和Littman[2]使用了点互信息(PMI-IR)方法,该方法利用了特定搜索引擎提供的"NEAR"操作来估计词汇与具有强烈倾向意义的种子词集合的关联程度,以此作为计算该词倾向性的依据。

当词语倾向性被计算出来后,文本的褒贬通常采用基于该倾向性词典的某种统计计算模型来判断。Turney[3]利用抽取出来的短语的倾向性来判断文本的倾向,在处理通用语料时可以达到74%;Pang[4]利用人工标注文本倾向性的训练语料,采用Bayesian、SVMs、最大熵学习分类器对电影评论进行分类研究,并证明了SVMs是最好的方法。然而,几乎所有的研究方法都是针对英文词汇的,对于中文的词语情感倾向性方法的研究还不是很多,文本倾向性分析技术还不够完善。

基于机器学习的情感分类方法利用训练语料对分类器进行训练,然后对测试语料进行实验,通过对特定领域中的语料进行训练来完成倾向性分析,但是这种分析方法在处理文本时缺乏语义信息。

基于语义理解的方法是利用词语相似度计算词语与褒义和贬义基准词的距离而得到词语的情感值[6],但是该方法没有考虑词语本身所具有的情感值。本文提出一种改进的基于语义理解的文本情感分类方法,用以更好地判别文本倾向性。

1 词语情感值计算

基于知网的语义相似度的方法反映的是词语语义的相似程度[5],亦即两个词语在不同上下文环境中在词语替换的情况下不改变文本句法语义结构的程度。因此,可以利用词语的语义相似度概念来计算词语的情感值。

1.1 基于HowNet的语义相似度的计算方法

到稿日期:2009-07-30 返修日期:2009-10-20 本文受国家自然科学基金(60773167),国家十一五科技支撑计划课题(2006BAK11B03),973国家重点基础研究发展计划(2007CB310804),教育部/国家外国专家局高等学校学科创新引智计划(B07042)资助。

闻 彬 主要研究方向为自然语言处理,E-mail:vanbrian@163.com;何婷婷 教授,博士生导师,主要研究方向为自然语言处理。

HowNet 中若词语有多种表达含义，则词语有多个义项，每个义项又由多个义原组成，那么词语的语义相似度计算实际上是义原的相似度计算。

对于两个词语 W_1 和 W_2，假设词语 W_1 有 n 个义项 $Y_1, Y_2, \cdots, Y_n$，词语 W_2 有 m 个义项 $Z_1, Z_2, \cdots, Z_m$，则词语的相似度计算如下：

$$\mathrm{Sim}(W_1, W_2) = \max_{i=1,\cdots,n, j=1,\cdots,m} \mathrm{Sim}(Y_i, Z_j) \quad (1)$$

将词语相似度的计算转换成概念之间的相似度计算。

1.2 义原相似度计算

在 HowNet 中概念由义原表示，所以概念相似度计算的前提是义原相似度计算。

由于所有义原根据上下位关系构成了一个树状的义原层次体系，因此可以使用式(2)计算两个义原之间的语义距离。

$$\mathrm{Sim}(p_1, p_2) = \frac{\alpha}{d+\alpha} \quad (2)$$

式中，p_1 和 p_2 表示两个义原，d 是 p_1 和 p_2 在义原层次体系中的路径距离，α 是一个可调节的参数。

1.3 概念情感相似度计算

在进行概念相似度计算时，将概念分成 4 个部分："第一基本义原"，"其他基本义原"，"关系义原"和"符号义原"。但是基于以上 4 种义原的相似度计算只考虑了词语替换的程度，没有考虑词语的情感语义。概念 S_1，S_2 之间的相似度计算如下：

$$\mathrm{Sim}(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} \mathrm{Sim}_j(S_1, S_2) \quad (3)$$

本文在计算情感相似度时引入了情感义原 $\mathrm{Sem}(Y, Z)$，改进后的概念 Y, Z 情感相似度具体计算如下：

$$\mathrm{Sim}(Y, Z) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} (\delta_1 \mathrm{Sim}_j(Y, Z) + \delta_2 \mathrm{Sem}(Y, Z)) \quad (4)$$

式中，$\mathrm{Sim}_1(Y, Z)$ 表示概念 Y 与 Z 的"第一基本义原"的相似度；$\mathrm{Sim}_2(Y, Z)$ 表示概念 Y 与 Z 的"其他基本义原"的相似度；$\mathrm{Sim}_3(Y, Z)$ 表示概念 Y 与 Z 的"关系义原"的相似度；$\mathrm{Sim}_4(Y, Z)$ 表示概念 Y 与 Z 的"符号义原"的相似度；$\mathrm{Sem}(Y, Z)$ 表示概念 Y 与 Z 的"情感义原"的相似度。当词语间的情感义原相同时，该值被设定为 1；当词语间的情感义原不相同或不存在时，该值被设定为 0。

β_i 是可调节的参数，其中 $i=1, \cdots, 4$，$\beta_1+\beta_2+\beta_3+\beta_4=1$，从 Sim_1 到 Sim_4 对于总体相似度所起的作用依次递减，因此设置的参数满足 $\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \beta_4$。

δ_1，δ_2 用来设置知网义原与情感义原的权重，$\delta_1+\delta_2=1$。

1.4 基于概念情感相似度的词语情感语义值

计算出概念情感相似度之后，需要计算词语的情感值。本文从 HowNet 中人工挑选出褒贬基准词集合 P_i 和 N_j。利用式(5)计算词语的情感值。

$$\mathrm{Sensibility}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Sim}(w, P_i) - \frac{1}{m}\sum_{j=1}^{m} \mathrm{Sim}(w, N_j) \quad (5)$$

式中，$\mathrm{Sensibility}(w)$ 表示词语 W 的情感值；$\mathrm{Sim}(w, P_i)$ 表示词语 w 与褒义基准词集合 P_i 的相似性；$\mathrm{Sim}(w, N_j)$ 表示词语 w 与贬义基准词集合 N_j 的相似性。

2 基于语义理解的文本情感分类方法

2.1 基于语义的文本情感分类

文本情感分类是根据文本所表达内容的情感倾向性将文本分成支持类、中立类和反对类。文本情感分类是未来多角度、立体型文本分类的一个重要研究方面，具有广泛的应用前景。基于语义理解的方法首先对文本进行预处理，得到标注词性的文本；然后利用对文本特定模式的抽取，得到文本的固定词性搭配；再结合词语情感值计算方法计算抽取出来的词语的情感值；最后通过文本情感计算获取文本的情感倾向性。基于上述过程的文本情感分类系统基本框架如图 1 所示。

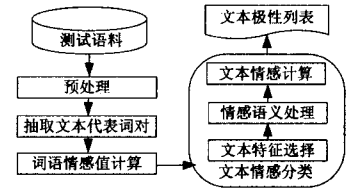

图 1 文本情感分类系统

2.2 文本特征选择

在文本中，评论者对于所发表的评论性文本，可能表达了"支持"、"中立"或"反对"的态度，但发表评论的目的必然是"推荐"或"不推荐"。因此，本文只对那些可能带有情感的词语进行抽取，比如形容词、名词和动词。为了更好地获取词语的语义倾向，需要建立如下统一的抽取规则。

规则 1：形容词＋名词

规则 2：副词＋形容词

规则 3：形容词＋形容词

规则 4：名词＋形容词

规则 5：副词＋动词

2.3 情感语义处理

在语义层次上，直接抽取出的情感词的倾向性可能不足以支持文本情感的准确表达。例如，"不理想"和"非常理想"，其中的情感词"理想"本来是褒义词，但是由于否定副词"不"和程度副词"非常"的出现，整个句子的语义在不同程度上都有所改变。因此，副词的出现可以对句子的倾向性加强、减弱甚至反向。研究副词及其出现规律对文本的情感倾向性显得尤为重要。

否定副词在文本中占有极其重要的地位，它的出现可以直接改变原有语句的情感倾向性。例如，"不/d 满意/v"、"不/d 差/a"等等。"满意"和"差"有其对应的倾向性，但由于否定副词的出现就使其表达了与原情感倾向性相反的含义。

本文从 HowNet 中人工抽取出 22 个否定副词，如表 1 所列，并从语料中抽取出否定规则。

表 1 否定副词表

否定副词
并非、不、不对、不再、不曾、不至于、从不、毫不、毫无、决非、绝非、没、没有、尚未、未、未必、未尝、未曾、永不、不大、不太、不很

对于词语的情感值，式(5)已经给出了计算方法，然而，当句子中出现了满足规则的否定副词时，则利用式(6)对词组的情感词极性进行调整。

$$SO(phrase) = (-1)^n * \mathrm{Sensibility}(w) \quad (6)$$

式中，$SO(phrase)$ 为抽取出的词组的情感值；n 为满足否定规则时对于词语 w 而言否定副词的出现次数。

除了否定副词之外，程度副词对句子的倾向性也起着不同程度的影响，可以对句子的情感有加强或者减弱的作用。例如：

"这款手机质量非常糟糕"

“X61 本本的性能特别优秀”

“糟糕”和“优秀”是情感词，并且在它们之前都出现了程度副词“非常”和“特别”，因此表达的情感强度比没有出现程度副词时更为强烈。本文抽取出 59 个程度副词并将其分成 7 类[7]，分别赋予不同的强度，从 1.5 到 0.8 倍，具体设置如表 2 所列。

表 2　程度副词赋值表

程度副词	1.5	最、最为、极、极为、极其、极度、极端
	1.4	太、至、至为、顶、过、过于、过分、分外、万分、何等
	1.3	很、挺、怪、老、非常、特别、相当、十分、甚、甚为、异常、深为、蛮、满、够、多、多么、殊、何其、尤其、无比、尤为
	1.2	不甚、不胜、好、好不、颇、颇为、大、大为
	1.1	稍稍，稍微，稍许，略微，略为，多少
	0.9	较、比较、较为、还
	0.8	有点、有些

利用式(6)计算程度副词结合情感词得到的情感值为

$$SO(Phrase)=value_{adv}*\text{Sensibility}(w) \tag{6}$$

式中，$SO(phrase)$表示抽取出的 $phrase$ 的情感值；Sensibility(w)表示情感词 w 的情感值；$value_{adv}$ 表示程度副词 adv 的强度值。

2.4　文本情感计算

计算出抽取的词对的情感值后，通过进行综合统计来判定文本的情感倾向性，即对所有词对情感值进行累加来得到文本的情感值。文本倾向性判定计算方法如式(7)所示。

$$orientation(d)=\sum_{i=1}^{n}SO(phrase_i) \tag{7}$$

式中，$phrase$ 为从文本中按固定模式抽取出来的词对。设定阈值 k，当 $orientation(d)\geqslant k$ 时，判定该文本为“推荐”文本；当 $orientation(d)<k$ 时，判定该文本为“不推荐”文本。

实验步骤如下：

Step1　对文档进行预处理，包括分词、词性标注并删除停用词；

Step2　利用开发的抽取工具对文本进行固定模式词语抽取，得到相应的词对；

Step3　利用本文提出的情感值计算方法计算抽取出来的词对的情感值；

Step4　利用文本分类方法判定文本情感倾向性。

3　实验分析

本文从网上下载电子产品领域的评论文档 435 篇，包括照相机、笔记本和手机 3 个领域，对挑选的语料进行人工褒贬倾向性判定，人工分类后的语料如表 3 所列。

表 3　语料类型

语料类型	倾向性		总数
	褒	贬	
照相机	83	63	146
笔记本	44	46	90
手机	102	97	199

对于知网中的词语，人工进行“desired/良”、“undesired/莠”标注，得到后词语词性数据如表 4 所列。

表 4　良莠人工标注结果

极性	词性				总数
	形容词	名词	动词	副词	
良	842	222	188	12	1264
莠	629	321	374	0	1324

利用本文提出的基于语义理解的文本情感分类方法对上述语料进行实验，并调整式(3)中的 δ_1，δ_2 的值。利用准确率作为实验结果的评估标准，实验数据如表 5 所列。

表 5　实验结果

δ_1，δ_2	k	照相机		照相机均值	k	笔记本		笔记本均值	k	手机		手机均值
		贬	褒			贬	褒			贬	褒	
0.2，0.8	38	0.873	0.771	0.822	54	0.913	0.727	0.820	52	0.969	0.745	0.857
0.3，0.7	32	0.873	0.771	0.822	48	0.934	0.727	0.831	44	0.959	0.765	0.862
0.4，0.6	28	0.873	0.771	0.822	41	0.934	0.727	0.831	38	0.969	0.765	0.867
0.5，0.5	22	0.873	0.783	0.828	34	0.913	0.727	0.820	31	0.969	0.765	0.867
0.6，0.4	18	0.873	0.783	0.828	25	0.870	0.772	0.821	24	0.959	0.775	0.867
0.7，0.3	14	0.889	0.783	0.836	18	0.870	0.772	0.821	17	0.948	0.775	0.861
0.8，0.2	8	0.873	0.783	0.828	11	0.870	0.795	0.833	10	0.928	0.784	0.856

当 δ_1，δ_2 设置成不同的值时，文本情感分类的准确率有不同程度的差异。对于每一次确定的 δ_1，δ_2 值，文本计算中的 k 都有唯一的值使得最后的判定结果最优。表 5 记录了在设定的 δ_1，δ_2 下，当 k 取何值时 3 类文档分别判定的准确率。从实验结果来看，随着 δ_1，δ_2 值不断调整，最高准确率基本趋于稳定，没有较大的波动，并且相对于传统的基于语义理解的方法有明显的提高，针对本文涉及到的 3 个领域的文本准确率最高分别可达 0.836，0.833 和 0.867。

在表 5 的基础上，表 6 重点考虑了 δ_1，δ_2 的值，并通过综合 3 类文档来计算判定精度。从实验结果可以发现，当 δ_1，δ_2 设定为 0.7，0.3 时，平均准确率达到了最高的 0.8393。结合表 5 与表 6 来分析实验数据，当 $\delta_1=0.7$，$\delta_2=0.3$ 时，对应的 k 值没有太大的变化幅度，其取值范围在 11 到 20 之间。因此，本文假设当 $\delta_1=0.7$，$\delta_2=0.3$ 时此区域可以保证各文档都能有较好的实验结果。表 7 考虑了 k 值在该条件下的判定影响程度。

表 6　δ_1，δ_2 对平均准确率的影响

δ_1，δ_2	均值			
	照相机	笔记本	手机	综合
0.2，0.8	0.822	0.820	0.857	0.8330
0.3，0.7	0.822	0.831	0.862	0.8383
0.4，0.6	0.822	0.831	0.867	0.8400
0.5，0.5	0.828	0.820	0.867	0.8383
0.6，0.4	0.828	0.821	0.867	0.8387
0.7，0.3	0.836	0.821	0.861	0.8393
0.8，0.2	0.828	0.833	0.856	0.8390

表 7　k 值的设置

均值				k
照相机	笔记本	手机	综合	
0.799	0.803	0.823	0.8083	11
0.812	0.802	0.834	0.8160	12
0.828	0.802	0.850	0.8267	13
0.836	0.791	0.860	0.8290	14

0.812	0.824	0.846	0.8273	15
0.800	0.811	0.851	0.8207	16
0.802	0.789	0.861	0.8173	17
0.798	0.821	0.852	0.8237	18
0.769	0.809	0.847	0.8083	19
0.773	0.820	0.837	0.8100	20

从表7中可以看到，当 $k\in[11,20]$ 时，实验结果都可以达到80%以上。因此，可以对 k 取该区域内的任意一个随机数来降低程序计算的复杂度，在简化实验过程的同时又不会降低判定准确率。

综合以上实验分析，当 $\delta_1=0.7,\delta_2=0.3$ 时，k 在闭区间[11,20]内取任意值，文本的情感分类准确率可以达到比较理想的结果，可以很有效地判别文本的情感倾向。

结束语 本文改进了词语相似度计算方法，提出了一种改进的基于语义理解的文本情感分类方法来判定文本的情感倾向性。通过实验分析，证明了该方法的有效性。

然而，由于没有公认统一的语料，本文的语料都是作者从网站中筛选出来的，语料的代表性还需要进一步研究。同时，在文本预处理之前，人工挑选出了情感强度比较大的句子作为实验语料，因此在文本情感值计算中，没有考虑句子的主客观性以及词语的语义消歧，这将在今后的研究中进一步细化。

参 考 文 献

[1] Hatzivassiloglou V, McKeown K R. Predicting the Semantic Orientation of Adjectives [A]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL[C]. 1997:174-181

[2] Peter T, Michael L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346

[3] Peter D T. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceeding of the Association for Computational Linguistics 40th Anniversary Meeting. New Brunswick, N. J., 2002

[4] Pang Bo, Lee Lillian, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002: 79-86

[5] 刘群，李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 台北，2002

[6] 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)：14-20

[7] 徐林宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007，21(1)：96-100

(上接第260页)

	σ(%)	0.077	**0.019**	0.019		σ(%)	0.059	0	0.059		σ(%)	0.632	0.415	**0.339**
	Nbest	2	2	**4**		Nbest	10	10	10		Nbest	0	0	**1**
	MTL	6217.5656	**6153.3945**	6199.6752		MTL	492.5287	492.3238	**491.7668**		MTL	2651.0134	**2641.2155**	2662.7500
*Ch130	MITL	6110.7219	**6110.7219**	6110.7219	*Gr202	MITL	**487.9470**	488.4269	488.9267	A280	MITL	2596.3825	2599.3292	**2586.7696**
6110.9	METL	6142.9789	**6127.5864**	6134.9483	549.99	METL	490.1223	**489.8160**	490.1981	2586.8	METL	2618.0728	2615.8590	**2609.581**
	MEI	291	304.7	**235.4**		MEI	**426.7**	471.5	456.8		MEI	**593.9**	781.2	781.2
	T(s)	**375.05**	499.38	584.06		T(s)	**703.39**	1163.45	1615.38		T(s)	**1601**	3085.66	3856.8
	σ(%)	0.526	**0.274**	0.395		σ(%)	0	0	0		σ(%)	1.209	1.123	**0.88**

结束语 本文将分层、局部最优免疫优势等概念应用于TSP问题，构造了一种基于抗体的HLOICSA算法，提高了人工免疫算法求解旅行商问题的效率。本算法在一定子种群规模情况下，通过对多个子种群进行局部最优免疫优势、克隆选择等操作，避免了抗体群被少数亲和力最高的抗体占满，增强了优秀抗体实现亲和力成熟的机会。又通过基于信息熵的抗体多样性改善操作，提高抗体群分布的多样性。低层操作为高层遗传算法产生更多种类的优良模式，通过高层遗传算法的操作，新子种群可以获得包含不同种类的优良模式的新抗体，从而为它们提供了更加平等的竞争机会。在保持优秀个体进化稳定性的同时，避免单个子种群进化过程中出现过早收敛现象，在深度搜索和广度寻优之间取得了平衡。

仿真试验表明，本文提出的HLOICSA算法在求解城市规模在700个城市以下(MATLAB在求解大规模TSP问题时存在计算时间过长的缺点，所以没有对700个以上城市进行验证)的TSP问题时非常有效，与NGA，IDIA，LOICSA相比，效率更高，全局搜索能力更强，收敛速度更快。通过子种群规模和克隆扩增算子中比例系数对HLOICSA性能影响的分析也表明HLOICSA具备良好的全局收敛可靠性以及较快的收敛速度。

参 考 文 献

[1] Michalewicz Z, Fogel D B. How to solve It: Modern Heuristick [M]. Berlin Heidelberg: Springer-Verlag, 2000

[2] Merz P, Freisleben B. Genetic local search for the TSP: New results [C]// Proc. of 1997 IEEE Int Conf on Evolutionary Computation. IEEE Neural Network Council, Evolutionary Programming Society, IEEE, 1997: 159-163

[3] de Castro L N, Von Zuben F J. The Clonal Selection Algorithm with Immune Systems and Their Applications[M]. CA: Morgan Kaufman Publishers, 2000

[4] Dasgupta D, Forrest S. Artificial immune systems in industrial applications[C] // Proc. of the Second Int Conf on IPMM'99. Honolulu, 1999, 1: 257-267

[5] 王磊，潘进，焦李成. 免疫规划[J]. 计算机学报，2000，23(8)：806-812

[6] 戚玉涛，刘芳，焦李成. 求解大规模TSP问题的自适应归约免疫算法[J]. 软件学报，2008，19(6)：1265-1273

[7] 焦李成，杜海峰，刘芳，等. 免疫优化计算，学习与识别[M]. 北京：科学出版社，2007：93-104，133-143

[8] 梁艳春，冯大鹏，周春光. 遗传算法求解旅行商问题时的基因判断保存[J]. 系统工程理论与实践，2000，4：7-12