文章编号：1003-0077(2009)05-0068-07

基于同义词的词汇情感倾向判别方法

王素格[1,3]，李德玉[2,3]，魏英杰[4]，宋晓雷[1]

(1. 山西大学 数学科学学院，山西 太原 030006；2. 山西大学 计算机与信息技术学院，山西 太原 030006；
3. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006；4. 科学出版社，北京 100717)

摘 要：词汇的情感倾向直接影响短语、句子、段落、篇章等更高层次语言粒度的情感倾向。对于基准词选取问题，该文提出了基于类别区分能力与情感词词表相结合的方法。考虑到词汇与其同义词很大程度上具有相同的情感倾向，我们提出了基于同义词的词汇情感倾向判别方法，这种方法一定程度上避免了数据稀疏问题。实验结果表明，基于同义词的词汇情感倾向判别方法优于仅采用目标词与基准词的词汇情感倾向判别方法。

关键词：计算机应用；中文信息处理；词汇情感倾向；基准词；关联强度；同义词

中图分类号：TP391　　**文献标识码**：A

A Synonyms Based Word Sentiment Orientation Discriminating

WANG Su-ge[1,3], LI De-yu[2,3], WEI Ying-jie[4], SONG Xiao-lei[1]

(1. School of Mathematics Science, Shanxi University, Taiyuan 030006, China;
2. School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China;
3. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China; 4. Science Press, Beijing 100717, China)

Abstract: The word sentiment orientation directly influences the sentiment orientation of higher level linguistic unit, such as the phrase, the sentence, the paragraph and the text. This paper proposes a paradigm word selection method based on the category distinguishing ability of a word and the sentiment word table. In consideration of that a word usually has the same sentiment orientation with its synonyms, we propose a method for word sentiment orientation discriminating based on synonyms. The method can avoid the data sparseness issue in a certain extent. The experiment results indicate that the proposed method is superior to the method based on the object word and paradigm words.

Key words: computer application; Chinese information processing; word sentiment orientation; paradigm word; relation intensity; synonym

1 引言

从语言学角度，语言粒度从小到大依次为语素、词、短语、句子、段落、篇章。文本的语义信息蕴含于各个层次的语言粒及语言粒的各种语法关系中。在计算语言学中，利用小粒度语言单元研究较大粒度语言单元是一种基于解析思想的常用方法。

作为最小语言粒度的语素，它是最小的音义结合体，其主要功能是构词。因此，在已有文本情感倾向分析的研究中，大多选择词作为基本的语言粒度，利用词的情感倾向确定搭配、句子、文本等的情感倾

收稿日期：2008-10-15　**定稿日期**：2009-05-04

基金项目：国家自然科学基金资助项目(60875040)；教育部科学技术研究重点基金(2007018)；教育部高等学校博士点基金(200801080006)；山西省自然科学基金资助项目(2007011042)；山西省重点实验室开放基金资助项目；山西高校科技研究开发项目(200611002)

作者简介：王素格(1964—)，女，博士，副教授，主要研究方向为自然语言理解、文本挖掘；李德玉(1965—)，男，教授，博导，主要研究方向为计算智能与数据挖掘；魏英杰(1982—)，男，硕士，主要研究方向为文本挖掘与自然语言处理。

向[1-3]。Turney[1]通过分析词汇上下文信息研究其情感倾向，采用 PMI-IR 方法，使用两个词汇作为种子来判断其他短语的语义倾向。之后，他们又在文献[2-3]中将单对种子扩展成多对种子，选取了正反面各 7 个词汇，分别采用 PMI-IR 和 LSA 两种方法来度量给定词汇与基准词的关联度，确定词汇的语义倾向，实验结果表明，PMI-IR 算法优于 LSA 方法。Dave 等[4]利用信息抽取技术从语料中产生特征(词汇)，以 Bayes 网络为工具分析各词汇与已标定情感类别文档之间的关系，进而计算各词汇的得分用于判定词汇的语义倾向。汉语词汇的情感倾向研究方面，香港城市大学的 Yuan 等[5]在 Turny 的工作基础上，对汉语极性词的自动获取进行了研究。复旦大学的朱嫣岚等[6]，提出了基于语义相似度和语义相关场的两种词汇语义倾向性计算方法，通过计算目标词汇与 HowNet 中已标注褒贬性词汇间的相似度，获取目标词汇的倾向性。大连理工大学的徐琳宏等[7]采用 HowNet 作为基准词，通过计算目标词与基准词的关联度，确定目标词汇的语义倾向。中国科学院自动化研究所的王根、赵军[8]提出了词语倾向性的极坐标方式，并使用了均衡化的互信息方法探讨了词语独立于上下文的自身倾向性。

上述文献[1-3,6-7]仅采用了目标词与基准词的关联强度来确定目标词的情感倾向，并没有考虑目标词与其同义词的关系，同时也没有对基准词的选择进行相关的研究。本文提出了基于类别区分能力与情感词表相结合的基准词选取方法，然后根据词汇与其同义词很大程度上具有相同的情感倾向的特点，提出了基于同义词的词汇情感倾向判别方法，该方法不仅考虑了目标词与基准词的关联强度，而且也考虑了目标词的同义词与基准词的关联强度。

2　词与词集间关联强度度量

(1) 词与词间的关联强度

点互信息(PointWise Mutual Information，PMI)是信息论中度量两个随机变量间统计依赖性的一种测度。利用 PMI 可以度量人们在使用某两个词的统计依赖性。设有两个词 $word1$ 和 $word2$，将两个词的使用看作两个随机变量，仍以 $word1$ 和 $word2$ 记之，进而有随机向量 $(word1, word2)$。在计算语言学中，常借用随机变量 $word1$ 和 $word2$ 的 PMI 值度量两个词 $word1$ 和 $word2$ 的统计依赖性[1,2]。

两个词 $word1$ 和 $word2$ 之间的点互信息 $PMI(word1, word2)$ 定义为：

$$PMI(word1, word2) = \log_2\left(\frac{P(word1, word2)}{P(word1)P(word2)}\right) \tag{1}$$

在实际应用中，公式(1)中的概率可以通过语料中两个词的同现信息进行估计。因此有下面的近似公式：

$$PMI(word1, word2) \approx \log_2\left(\frac{N \times hits(word1, word2)}{hits(word1) \times hits(word2)}\right) \tag{2}$$

这里，N 表示语料库中总的词次数，$hits(word1)$ 和 $hits(word2)$ 分别表示 $word1$ 和 $word2$ 在语料库中出现的次数，$hits(word1, word2)$ 表示 $word1$ 和 $word2$ 在语料库中限定观察范围的同现次数。

(2) 词与词集间关联强度

词与词集间关联强度可由词与词间关联强度来计算。设 $word$ 是一个词，$wordSet$ 是一个词集，定义词 $word$ 与词集 $wordSet$ 的关联强度如下：

$$PMI(word, wordSet) = \sum_{word' \in wordSet} \log_2\left(\frac{P(word, word')}{P(word)P(word')}\right) \tag{3}$$

3　基于同义词的词汇情感倾向判别

基准词集：基准词集是指褒贬义倾向非常明显、强烈、具有代表性的词汇所构成的集合。基准词集被分为褒义基准词集和贬义基准词集，分别记为 $Pwords$ 和 $Nwords$。

词的情感倾向强度：一个词的情感倾向强度可由该词与褒义基准词集和贬义基准词集的关联强度的差来计算，由公式(3)，设 $word$ 是一个词，则 $word$ 的情感倾向强度 $SO\_PMI(word)$ 为：

$$SO\_PMI(word) = \sum_{pword \in Pwords} PMI(word, pword) - \sum_{nword \in Nwords} PMI(word, nword) \tag{4}$$

再由公式(2)和(4)，得出

$$SO\_PMI(word) \approx \log_2\left(\frac{\prod_{pword \in Pwords} hits(word, pword) \prod_{nword \in Nwords} hits(nword)}{\prod_{pword \in Pwords} hits(pword) \prod_{nword \in Nwords} hits(word, nword)}\right) \tag{5}$$

一个词与褒义基准词集的关联强度越大，则该词倾向于褒义的程度就越大，反之，它与贬义基准词集的关联强度越大，则其倾向于贬义的程度就越大。词的情感倾向强度 $SO\_PMI(word)$ 刻画了一个词更倾向于褒义还是贬义的程度。

基于同义词的词汇情感倾向强度：设 $word$ 是一个词，$T=\{a_i\}_{i=1}^{n}$ 是词 $word$ 的同义词集合，为了区分词与其同义词对词汇情感倾向强度判断的贡献，将 α,β 作为权重，构造出如下计算 $word$ 的词汇情感倾向强度公式：

$$\begin{aligned} &New\_SO\_PMI(word)\\ &=\alpha \cdot SO\_PMI(word)\\ &\quad +\beta \cdot \sum_{i=1}^{n} SO\_PMI(a_i) \end{aligned} \tag{6}$$

这里 $\alpha+\beta=1,\alpha,\beta$ 分别表示目标词与其同义词的情感倾向强度对最后目标词的情感倾向强度的影响程度。特别地，当 $\alpha=1,\beta=0$ 时，$New\_SO\_PMI(word)=SO\_PMI(word)$，即为直接使用词的情感倾向强度。

由于写作的习惯不同，不同的作者在撰写评论时，会使用不同的词汇表达相同的意思。即使同一作者，在一篇评论中也常常为避免重复而使用同义词和近义词表达相同的意思。比如，真实语料中有下面两个例句：

(1) 新 POLO 的悬架经过调拨后，舒适性有所增强，配合舒适的座椅，那种冲过坑洼的颠簸感只是在踏板上能有清晰感觉。

(2) 新车强调驾乘乐趣，即保证宝马良好操纵性能的基础上加强乘坐的舒适性，着力营造良好的商务空间。

在这两个句子中“增强”和“加强”是同义词，在同义词词林中列出词条“增强”、“加强”、“提高”、“增高”、“增进”、“增长”、“滋长”、“如虎添翼”均为同义词。

就考察词的倾向性而言，从统计词的角度看，将一个词与其同义词或近义词按不同词对待，将会导致大量稀疏数据。

在自然语言处理中，数据稀疏一直是困扰人们的一大问题，单纯考察一个词与褒贬义基准词集的同现信息就会遇到数据稀疏问题。这里，我们提出的基于同义词的词汇情感倾向强度计算方法，利用一个词的同义词集来重新定义该词的情感倾向强度，在某种程度上弱化了数据稀疏问题。

词汇情感倾向类别确定：

对一个词 $word$，利用公式(6)可以计算其情感倾向强度，设 $\theta_1,\theta_2(\theta_1\geqslant\theta_2)$ 是两个实数，称为阈值，词 $word$ 的情感类别 $So(word)$ 可由判别公式(7)得到。

$$So(word)=\begin{cases} \text{褒义}, & New\_SO\_PMI(word)>\theta_1\\ \text{中性}, & \theta_2\leqslant New\_SO\_PMI(word)\leqslant\theta_1\\ \text{贬义}, & New\_SO\_PMI(word)<\theta_2 \end{cases} \tag{7}$$

4 基准词的选取方法

由第 3 节知，词汇的情感倾向强度计算需要基准词集，而基准词是指具有非常明显、褒贬义倾向的代表性词汇。为此，本文提出了基于词汇的类别区分能力与情感词表相结合的基准词选取方法。

情感词表主要借助 General Inquirer(GI)词典、《学生褒贬义词典》[9]、《知网》、《褒义词词典》[10]、《贬义词词典》[11]五种资源构建的中文情感词词表，记为 SWT。该词表共收录词条 15 886 个(正面 8 427 个，反面 7 459 个)，其中仅来源于一部词典的词条 11 682 个(正面为 6 129 个，反面为 5 553 个)。另有来源于多个词典的词条 4 204 个(正面为 2 298 个，反面为 1 906 个)。详细情况参见文献[12]。

词汇的类别区分能力的度量采用一种与文本长度无关的基于词频(频率)概率估计的 Fisher 准则函数计算方法[12]。

设正面文本有 m 篇，记为 $d_{P,i}(i=1,2,\cdots,m)$，$d_{P,i}$ 的总词次记为 $v_{P,i}$，特征项 t_k 在 $d_{P,i}$ 中出现的次数记为 $w_{P,i}(t_k)$。设反面文本有 n 篇，记为 $d_{N,j}(j=1,2,\cdots,n)$，$d_{N,j}$ 的总词次记为 $v_{N,j}$，特征项 t_k 在 $d_{N,j}$ 中出现的次数记为 $w_{N,j}(t_k)$。从 m 个正面文本中任取一篇，显然 $P(d_{P,i})=\frac{1}{m}$，现以 $\frac{w_{P,i}(t_k)}{v_{P,i}}$ 代替文本 $d_{P,i}$，则有随机变量 $X1$：$P(d_{P,i})=P\left(\frac{w_{P,i}(t_k)}{v_{P,i}}\right)=\frac{1}{m}(i=1,2,\cdots,m)$。对反面文本作类似的考虑，则有随机变量 $X2$：$P(d_{N,j})=P\left(\frac{w_{N,j}(t_k)}{v_{N,j}}\right)=\frac{1}{n}(j=1,2,\cdots,n)$。这样则有，

$$E(t_k\mid P)=E(X1)=\frac{1}{m}\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}},$$

$$E(t_k\mid N)=E(X2)=\frac{1}{n}\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}},$$

$$D(X1)=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{w_{P,i}(t_k)}{v_{P,i}}-E(X1)\right)^2,$$

$$D(X2)=\frac{1}{n}\sum_{j=1}^{m}\left(\frac{w_{N,j}(t_k)}{v_{N,j}}-E(X2)\right)^2$$

基于词频(频率)概率估计的 Fisher 准则函数计算有：

$$F(t_k)=\frac{(E(t_k\mid P)-E(t_k\mid N))^2}{D(t_k\mid P)+D(t_k\mid N)}$$

$$=\frac{m\times n\times\left(n\times\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}}-m\times\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2}{n^2\times\sum_{i=1}^{m}\left(m\times\frac{w_{P,i}(t_k)}{v_{P,i}}-\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}}\right)^2+m^2\times\sum_{j=1}^{n}\left(n\times\frac{w_{N,j}(t_k)}{v_{N,j}}-\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}}\right)^2}\quad(8)$$

基于词汇类别区分能力与情感词表相结合的基准词选取方法的具体步骤：

1) 利用公式(8)，计算语料库中名词、形容词和动词的类别区分能力，选出区分能力较强的词 M 个，获得词集 CWordSet。

2) 用词集 CWordSet 与情感词词表 SWT 做交集，得到词集 CBWordSet，并将 CBWordSet 中的词按照其在语料中出现的次数排序，得到的词集记为 DCBWordSet。

3) 根据事先设定的基准词个数 N，选取词集 DCBWordSet 中的正反类别中各前 N 个词，作为最终选定的褒贬基准词集 BWordSet。

5　实验结果与分析

为了测试本文提出方法的有效性，我们选用来自汽车点评网自建的语料。本语料收集了国内外 11 种品牌的轿车，评论时间集中于 2006 年 1 月至 2007 年 3 月间的部分评论文本，总计 1 006 篇约 100 万字，正面文本 578 篇，反面文本 428 篇。情感词表选用仅来源于一部词典 11 682 个词条，同义词词集采用张伟等人编纂的《学生褒贬义词典》[9]和哈尔滨工业大学信息检索研究室提供的《同义词词林扩展版》[13]两部词典。

评价指标采用标注精确率(P)、召回率(R)和 F 值以及正反面精确率(PP、NP)、正反面召回率(PR、NR)和正反面 F 值(PF、NF)。由于本文只对词汇的两种情感倾向性进行判别，因此总体的评价指标 $P=R=F$。

根据第 4 节中基准词选取步骤，选取 $M=4\,000$，$N=40$，得到褒贬基准词集如表 1、表 2 所示。

表 1　褒义基准词集

好	安全	不错	喜欢	加速	舒适	豪华	满意	爱	解决
风格	优势	保证	全新	实在	舒服	稳定	方便	品质	提升
乐趣	省油	先进	成功	漂亮	最好	保护	好车	值得	良好
满足	享受	出色	提高	适合	平稳	轻松	优点	完美	实用

表 2　贬义基准词集

碰撞	噪音	事故	毛病	不好	严重	下降	缺点	不够	死
不足	故障	缺陷	郁闷	撞击	断裂	失望	担心	倒	车祸
遗憾	怀疑	不行	变形	断	危险	震动	损失	噪声	麻烦
冲击	隐患	后悔	恐怕	粗糙	颠簸	造成	难看	不爽	伤害

测试词集选用语料中的词集与情感词汇词表交集的词汇，共有 2 958 个。采用两种方式进行实验。

1) 面向语料

为了验证本文方法的有效性以及与领域的相关性，实验选用的测试语料为 1 006 篇汽车评论，观察同现窗口长度设定为 24 个词位，词汇情感强度计算分别采用公式(5)和公式(6)，阈值的选取采用试验法。

实验 1：为了验证本文提出的情感基准词选取方法的优势，将本文选出的基准词与文献[6]列出的基准词进行了对比实验，基准词分别选出 40 对、前 10 对和前 5 对。测试结果见表 3。

由表 3 可知：

① 随着基准词数量的增加，词汇的情感倾向判别的精确率逐渐升高。

② 利用本文选择的基准词得到的词汇情感分类结果整体优于文献[6]提供基准词的结果。

表3 不同基准词对词汇情感倾向判别的影响

基准词	权值		40对		10对		5对	
	α	β	阈值	P/%	阈值	P/%	阈值	P/%
本文	1	0	−0.340 4	67.51	−0.552 6	67.00	−0.552 6	65.78
	0.5	0.5	−0.174 5	69.81	−0.182 3	68.73	−0.552 6	67.41
文献[6]	1	0	4.069 7	56.28	2.490 3	46.14	2.274 0	45.20
	0.5	0.5	3.361 8	61.16	2.514 2	55.10	2.274 0	54.77

③ 在两种基准词集下，采用基于同义词的词汇的情感倾向判别的精确率相比基于词的词汇的情感倾向判别的精确率有所提高。特别地，采用文献[6]中的基准词的提高幅度较大，当选用5对基准词时提高了9.57%，说明文献[6]中的基准词具有通用性。

综合上述结果说明，在特定领域中，若统计文献[6]中的基准词与其他词汇的同现次数时，将会出现大量的数据稀疏现象，若采用基于同义词的词汇的情感倾向判别，在一定程度上可以减少数据稀疏，并提高词汇的情感倾向判别精度。但总体上，采用文献[6]中的基准词得到词汇情感倾向判别的结果逊色于本文的方法，因此，对特定领域的情感倾向性判别，应选择面向领域的基准词集，避免使用通用基准词集。以下实验中的基准词集均选用表1和表2中的40对基准词。

实验2：由于我们采用的基于同义词的词汇的情感倾向判别方法，在一定程度上依赖于词的同义词，因此，采用了以下两种方法进行了对比实验。

方法A(基于同义词的词汇情感倾向判别)：对α和β，分别采用五组不同的值得到词的情感倾向；

方法B(直接使用同义词词典)：采用基准词的情感倾向和同义词词典，用于词的情感倾向判别。

上述两种方法得到实验结果见表4。

表4 采用方法A和方法B的词汇情感倾向判别结果

方法	权值		阈值	PP	PR	PF	NP	NR	NF	P
	α	β								
方法A	1	0	−0.340 4	72.09	85.48	78.21	48.14	28.94	36.15	67.51
	0.8	0.2	−0.084 9	73.16	88.21	79.98	54.67	30.51	39.18	69.88
	0.7	0.3	−0.139 7	73.04	88.35	79.97	54.55	30.00	38.71	69.81
	0.6	0.4	−0.19.88	73.10	88.60	80.11	55.08	30.00	38.84	69.98
	0.5	0.5	−0,17.45	72.89	88.75	80.04	54.69	29.15	38.03	69.81
方法B				98.70	7.53	14.00	92.31	3.83	7.54	6.36

实验3：为了进一步说明同义词在词汇情感倾向判别的作用，我们去掉没有同义词的词，仅仅对含有同义词的词采用方法A和方法B，重复实验2的过程，得到的实验结果见表5。

表5 对含有同义词的词采用方法A和方法B的词汇情感倾向判别结果

方法	权值		阈值	PP	PR	PF	NP	NR	NF	P
	α	β								
方法A	1	0	−0.808 2	75.39	87.38	80.94	49.51	30.27	37.57	70.80
	0.8	0.2	−0.928 7	77.36	94.54	85.09	70.78	32.34	44.40	76.49
	0.7	0.3	−0.709 5	77.30	94.66	85.11	71.05	32.05	44.17	76.49
	0.6	0.4	2.157 7	79.83	90.29	84.74	65.07	44.21	52.65	76.92
	0.5	0.5	4.290 5	80.00	90.29	84.83	65.37	44.81	53.17	77.09
方法B				98.70	18.45	31.08	92.31	10.68	19.15	16.19

由表 4 和表 5 可知：

① 方法 A 和方法 B 相比，后者得到的词汇情感倾向判别的精确率优于前者，而其他各项评价指标都明显劣于前者。说明仅仅直接使用同义词词典可以得到比较高的精确率，但却由于匹配的词汇较少，造成了较低的召回率。

② 将 $\beta=0$ 与 $0<\beta\leqslant 0.5$ 相比，后者得到的词汇情感倾向判别的各项评价指标明显优于前者。而将表 4 与表 5 相比，后者得到的词汇情感倾向判别的各项指标均优于前者，说明同义词在词汇情感倾向判别时确实发挥了作用，提高了词的情感倾向识别的的总体精度。

③ 对于 $0<\beta\leqslant 0.5$ 时，两表中四种情况的词汇情感倾向判别结果的总体精度（P）都相差不大，验证了我们在第 3 节中的最初设想，具有同义词的不同词语可以表达相同的语义信息。

④ 褒义词汇的精确率（PP）、召回率（PR）和 F 值（PF）均优于贬义词汇的精确率（NP）、召回率（NR）和 F 值（NF）。主要原因我们测试的 2 958 个词语中贬义词占的比例比较小，仅有 941 个。

2) 面向 Web

实验 4：在实验 1、实验 2 中，由于语料规模的限制，词汇的统计数据比较稀疏。为此，本实验进行了面向 Web 的实验测试。选用 Google 搜索引擎，将互联网页作为资源，Google 作为目前最成功的商业搜索引擎之一，索引的网页数量已超过 80 亿。由于本文的 PMI 需要进行批量查询，因此利用了 Google API。然而 Google API 的不足之处在于其返回的相关网页数量是一个估计值，这可能会给 PMI 计算模型引入一些噪音。但是从总体上看，几个查询之间返回的网页数量比例还是相对比较稳定的。

由于 Google 没有提供 NEAR 操作，所以观察两个词 *word*1 和 *word*2 的同现是将两个词作为查询词共同提交给 Google 进行检索，即将查询的窗口尺寸大小设定为整篇文档。

实验采用表 1 和表 2 中 40 对褒贬义基准词，当采用词的情感倾向强度时，精确率为70.97%，当采用同义词的词的情感倾向强度时，精确率为76.20%。由此结果可以看出：

① 基于同义词的词汇情感倾向强度方法的分类效果优于基于词的情感倾向强度的方法。再次验证了利用同义词集确实可以改善词的情感倾向识别。

② 与表 3 对比可知，两种词汇情感强度计算方法在面向 Web 的测试结果均优于面向语料的测试结果。说明语料规模对词汇情感倾向强度的计算有较大影响。

6　结束语

词汇作为构成短语、搭配、关联对、句子和文本的最基本的语言粒度，其情感倾向直接影响更高层次语言粒度的情感倾向。本文提出了基于类别区分能力的基准词选择方法，并根据词汇与其同义词具有相近的褒贬情感倾向的特点，提出了基于同义词的词汇情感倾向判别方法。本文所提出的方法，一方面，从文本情感分类的角度，利用词汇的情感倾向可以确定出短语、搭配、关联对、句子等语言粒度的情感倾向，最终确定出文本的情感倾向，另一方面，从情感词表构建的角度，也可以实现词表的动态更新。

感谢

感谢哈尔滨工业大学信息检索研究室为我们的研究提供了《同义词词林扩展版》，感谢董振东先生为我们的研究提供《知网》中的评价词汇和情感词汇。

参考文献：

[1] PETER D. Turney. 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 (ACL)//Philadelphia, PA, USA. 2002: 417-424.

[2] PETER D. Turney and MICHAEL L. Littman. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.

[3] PETER D. Turney and MICHAEL L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus [R]. Tech. Rep. EGB-1094, National Research Council Canada: 2002.

[4] DAVE K., LAWRENCE S., and PENNOCK D.. Mining the peanut gallery: opinion extraction and semantic classification of product reviews [C]//Proceedings of the 22nd International World Wide Web Conference. Budapest, Hungary: 2003.

[5] YUEN Raymond W. M., CHAN Terence Y. W., LAI Tom B. Y. et al. Morpheme-based derivation of bipolar semantic orientation of Chinese words [C]//Proc. Of the 20th International Conference on Computational Linguistics (COLING-2004), Geneva, Switzerland. 2004: 1008-1014.

[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，2006，21(1)：14-20.

[7] 徐琳宏，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.

[8] 王根，赵军. 中文褒贬义词语倾向性的分析[C]//第三届学生计算语言学研讨会论文集. 沈阳. 2006：81-85.

[9] 张伟，刘缙，郭先珍. 学生褒贬义词典[M]. 中国大百科全书出版社. 2004.

[10] 史继林，朱英贵. 褒义词词典[M]. 四川：四川辞书出版社. 2005.

[11] 杨玲，朱英贵. 贬义词词典[M]. 四川：四川辞书出版社. 2005.

[12] 王素格. 基于Web的评论文本的情感分类问题研究[D]. 博士论文. 上海：上海大学. 2008.

(上接第61页)

Meeting of the Association for Computational Linguistics, Hong Kong, China. 1999.

[23] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, and H. Isahara. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules [C]//Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China. 2000.

[24] T. Kudo, and Y. Matsumoto. Chunking with Support Vector Machines [C]//Proceedings of Second Meeting of North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA. 2001.

[25] Z. P. Jiang, J. Li, H. T. Ng. Semantic Argument Classification Exploiting Argument Interdependence [C]//Proceedings of 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005: 1067-1072.

[26] H. T. Ng and J. K. Low. Chinese Part-Of-Speech Tagging: One-At-A-Time Or All-At-Once? Word-Based Or Character-Based? [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain. 2004.

[27] H. Duan, X. Bai, B. Chang, S. Yu. Chinese word segmentation at Peking University[C]//Proceedings of the second SIGHAN workshop on Chinese language processing. Sapporo, Japan, 2003: 152-155.

[28] V. Punyakanok, D. Roth, W. Yih. The importance of syntactic parsing and inference in semantic role labeling[J]. Computational Linguistics, 2008, 34(2): 257-287.

基于同义词的词汇情感倾向判别方法

作者： 王素格， 李德玉， 魏英杰， 宋晓雷， WANG Su-ge， LI De-yu， WEI Ying-jie， SONG Xiao-lei

作者单位： 王素格,WANG Su-ge(山西大学,数学科学学院,山西,太原,030006;山西大学,计算智能与中文信息处理教育部重点实验室,山西,太原,030006)， 李德玉,LI De-yu(山西大学,计算机与信息技术学院,山西,太原,030006;山西大学,计算智能与中文信息处理教育部重点实验室,山西,太原,030006)， 魏英杰,WEI Ying-jie(科学出版社,北京,100717)， 宋晓雷,SONG Xiao-lei(山西大学,数学科学学院,山西,太原,030006)

刊名： 中文信息学报 ISTIC PKU

英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING

年，卷(期)： 2009,23(5)

参考文献(12条)

1. PETER D Turney;MICHAEL L Littman Unsupervised learning of semantic orientation from a hundred-billion-word corpus.[Tech.Rep.EGB-1094] 2002

2. PETER D Turney;MICHAEL L Littman Measuring praise and criticism:inference of semantic orientation from association[外文期刊] 2003(04)

3. PETER D Turney Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews 2002

4. 王素格 基于Web的评论文本的情感分类问题研究 2008

5. 杨玲;朱英贵 贬义词词典 2005

6. 史继林;朱英贵 褒义词词典 2005

7. 张伟;刘缙;郭先珍 学生褒贬义词典 2004

8. 王根;赵军 中文褒贬义词语倾向性的分析 2006

9. 徐琳宏;林鸿飞;杨志豪 基于语义理解的文本倾向性识别机制[期刊论文]-中文信息学报 2007(1)

10. 朱嫣岚;闵锦;周雅倩 基于HowNet的词汇语义倾向计算[期刊论文]-中文信息学报 2006(01)

11. YUEN Raymond W M;CHAN Terence Y W;LAI Tom B Y Morpheme-based derivation of bipolar semantic orientation of Chinese words 2004

12. DAVE K;LAWRENCE S;PENNOCK D Mining the peanut gallery.,opinion extraction and semantic classification of product reviews 2003

引证文献(2条)

1. 彭学仕.孙春华 面向倾向性分析的基于词聚类的基准词选择方法[期刊论文]-计算机应用研究 2011(1)

2. 彭学仕.孙春华 面向倾向性分析的基于词聚类的基准词选择方法[期刊论文]-计算机应用研究 2011(1)