第 38 卷 第 15 期
Vol.38 No.15

计 算 机 工 程
Computer Engineering

2012 年 8 月
August 2012

• 人工智能及识别技术 • 文章编号：1000—3428(2012)15—0187—03 文献标识码：A 中图分类号：TP391

基于 HowNet 和 PMI 的词语情感极性计算

王振宇[a]，吴泽衡[b]，胡方涛[a]

(华南理工大学 a. 软件学院；b. 计算机科学与工程学院，广州 510006)

摘 要：基于语料库的点互信息(PMI)计算方法依赖于语料库的完善性，基于 HowNet 的计算方法则依赖于知网相似度计算的准确性。为克服 2 种方法的局限性，提出一种 HowNet 和 PMI 相融合的词语极性计算方法，利用知网进行同义词扩展，降低情感词在语料库中出现频率低所带来的问题。实验结果表明，该方法的微平均和宏平均性能比传统方法提升约 5%。

关键词：情感分析；点互信息；知网；同义词扩展；相似度

Words Sentiment Polarity Calculation Based on HowNet and PMI

WANG Zhen-yu[a], WU Ze-heng[b], HU Fang-tao[a]

(a. School of Software; b. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

【Abstract】The polarity calculation of word level is the basis of sentiment analysis of sentence level and discourse level. The traditional calculation methods based on Point Mutual Information(PMI) or HowNet have their own defects: methods of PMI depend on the perfection of the corpus, and methods of HowNet depend on accuracy of the similarity calculation based on HowNet. In order to improve these deficiencies, an improved method for calculating the polarity of words is proposed, combining HowNet with PMI. First of all, HowNet is used to expand the synonyms of the emotional words in order to reduce the impact of some emotional words which have low frequency in the corpus, and then, according to the similarity calculation based on HowNet, it integrates the similarity based on HowNet with that of PMI. Experimental results show the new method increases micro average and macro average by 5% compared with traditional methods.

【Key words】sentiment analysis; Point Mutual Information(PMI); HowNet; synonym expansion; similarity

DOI: 10.3969/j.issn.1000-3428.2012.15.052

1 概述

词语级的情感极性分析是句子级和篇章级的情感极性分析的基础和前提，它包括 2 个方面的含义：提取出可能具有情感倾向的候选词；对该候选词进行分析，判断其倾向性及极性强度。中文文本的情感词一般以形容词、动词、名词、副词为主。词语的情感极性计算主要有 2 种方法：基于词典的方法和基于语料库的方法[1]。

基于词典的方法主要是利用词典中词语之间的相互联系来挖掘情感词。英文中常用的方法是计算情感词与种子词在 Wordnet[2]中的关联程度来识别出情感词；文献[3]采用词典中词语的注释信息来完成情感词识别的极性计算。中文中常用的词典是知网(HowNet)[4]，文献[5]提出一种基于 HowNet 的词汇语义倾向计算方法，取得了较好的效果。

基于语料库的方法主要是搜集一个足够大的语料库，利用其统计特征计算词语极性。常用方法是定义一些种子词，根据新词与种子词的紧密程度，判断新词的情感倾向和极性强度。文献[6]提出一种点互信息(Point Mutual Information, PMI)的方法计算某个词语的情感极性，该方法通过计算情感词与基准词在语料库中的共现概率计算新词的情感极性。

本文对中文词语情感极性计算方法并作简要介绍与分析，提出一种改进的词汇语义倾向计算方法，该方法首先通过知网来扩展同义词，然后将知网相似度与 PMI 计算进行融合，并进行实验验证[7]。

2 中文词语情感极性计算方法

中文词语常用的情感极性计算方法是基于 PMI 的计算方法和基于知网的计算方法[8]。基于 PMI 的词语情感极性计算方法首先选取一些基准词，这些基准词有代表褒义的，也有代表贬义的，通过计算新词与这些基准词在语料库中的共现概率，确定新词的褒贬义倾向及强度。

假定基准褒义词为 $WordSet_1=\{commendatory_1, commendatory_2,\cdots, commendatory_n\}$，贬义词为 $WordSet_2=\{derogatory_1, derogatory_2,\cdots,derogatory_n\}$，则对于某个词 *Word*，基于 PMI 的词语极性 *SO_PMI*(*Word*)的计算如式(1)所示：

$$SO\_PMI(Word)=\sum_{i=1}^{n}PMI(Word,commendatory_i)-\sum_{i=1}^{n}PMI(Word,derogatory_i) \quad (1)$$

PMI 的计算公式为：

$$PMI(Word_1,Word_2)=\mathrm{lb}\left(\frac{P(Word_1\ \&\ Word_2)}{P(Word_1)P(Word_2)}\right)$$

其中，$P(Word_1)$表示 $Word_1$ 在语料库中独立出现的概率；$P(Word_2)$表示 $Word_2$ 在语料库中独立出现的概率；$P(Word_1\&$

基金项目：广东省科技计划基金资助项目“基于情感极性分析的互联网敏感信息监控系统项目号”(2010B010600017)

作者简介：王振宇(1967－)，男，教授、博士、博士生导师，主研方向：中文信息处理，分布式系统；吴泽衡、胡方涛，硕士研究生

收稿日期：2011-08-09 **修回日期：**2011-11-12 **E-mail：**147188611@qq.com

$Word_2$)表示 $Word_1$ 和 $Word_2$ 同时出现的概率。

基于知网相似度的词语情感极性计算方法是一种基于电子词典的方法，它与上一种方法类似，也是选取一些基准词，然后根据情感词与基准词的紧密程度对情感词进行计算。不同的是，上一种方法的紧密程度是通过语料库中词语的共现情况来计算的，而这种方法是通过语义词典知网来计算的。

对于某个词 *Word*，基于知网相似度的词语极性 *SO_Hownet*(*Word*)的计算如下：

$$SO\_Hownet(Word)=\sum_{i=1}^{n} Sim(Word,commendatory_i)-\sum_{i=1}^{n} Sim(Word,derogatory_i) \quad (2)$$

其中，$Sim(Word,commendatory_i)$、$Sim(Word,derogatory_i)$分别表示知网的相似度计算，即 *Word* 与其褒义词 $commendatory_i$、贬义词 $derogatory_i$ 的相似度。

以上 2 种方法都存在缺点。基于 PMI 的方法的主要缺点是：PMI 的计算依赖于语料库，如果某些情感词在语料库中出现的概率较低，这种方法将无法得到情感词的正确极性。例如，“驰誉”这个词表达的意思是驰名、著名，是一个正向评价，但是其在语料库中出现的次数很少，在本文的语料库中，它出现的次数甚至是 0，导致其 PMI 的计算结果为 0，将其判定为中性词。这样的词汇并不少，因此，它们将导致算法的性能大幅下降。

基于知网的方法的缺点是：计算情感词与基准词之间的紧密程度主要通过知网相似度来计算，而知网相似度的计算又是通过义原在义原树中的距离来度量；由于知网由人工整理，不可能对所有词都十分完善，某些词尽管非常相似，但它们的义原在义原树中距离却很远，因此相似度很低。通过观察发现，如果 2 个词通过知网相似度计算有较高的相似度，一般来说它们确实是相似的；但如果 2 个词通过知网相似度计算后有较低的相似度，则它们不一定是不相似的。例如，靓丽和美丽是 2 个很相似的词，在辞海中，靓丽的解释即为漂亮、美丽；而通过知网相似度计算，靓丽与美丽的相似度计算结果只有 0.16，相似程度非常低。这种情况导致词语的极性计算不准确。

3 HowNet 与 PMI 相融合的词语极性计算

3.1 基于知网的同义词扩展

中文词语之间存在着大量的联系，最直接的联系就是某些词语之间存在的同义关系，例如，“驰誉”和“驰名”这 2 个词，表达的意义非常相近，大多数情况下可替换使用。如果能利用这种同义关系，无疑会提高计算的准确率。如上节介绍，“驰誉”这个词在语料库中出现的概率很低，在计算“驰誉”与基准词的共现概率时，由于 *P*(驰誉&基准词)和 *P*(驰誉)的值都非常小，甚至是 0，导致 PMI 的值非常小，因此最后可能将“驰誉”判定为中性词。由于“驰名”这个词在语料库中出现得非常频繁，因此如果能够扩展“驰誉”的同义词“驰名”，将其加入计算，那么就可以通过计算“驰名”的 PMI 值得出“驰誉”的 PMI 值。

基于以上分析，算法的第一步对同义词进行扩展，对于每个需要计算情感极性的词汇，通过知网语义词典来进行扩展。根据知网对概念的描述，察看“驰誉”和“驰名”在知网中的描述。

驰誉的概念如下：

NO.=020728
W_C=驰誉
G_C=adj [chi2 yv4]
S_C=PlusSentiment|正面评价
E_C=
W_E=famous
G_E=adj
S_E=PlusSentiment|正面评价
E_E=
DEF={famous|著名}

驰名的概念如下：

NO.=020714
W_C=驰名
G_C=adj [chi2 ming2]
S_C=PlusSentiment|正面评价
E_C=
W_E=famous
G_E=adj
S_E=PlusSentiment|正面评价
E_E=
DEF={famous|著名}

其中，NO.是概念的编号；W_C 是中文词语；G_C 是中文词性和音标；E_C 是中文的例子；W_E 是英文词语；G_E 是英文词性；E_E 是英文例子；DEF 是用知网描述语言 KDML 描述的概念。从中可见，它们的 W_E 与 DEF 是相同的，通过观察大量词语，本文给出基于知网的同义词扩展方法：如果 2 个概念的 W_E 和 DEF 相同，那么它们就是同义词。

3.2 算法描述

扩展同义词之后，将它们加入到词语倾向计算中。对于 2 个词汇，如果它们的知网相似度计算结果很高，那么它们是相似的。本文认为，如果待计算词汇(包括其同义词)与预先定义好的某个基准词之间的相似度很高，则说明它们在大部分情况下可以相互替换，其极性倾向一致；而极性强度则可通过它们之间的相似度及基准词的极性强度得到。如果待计算词汇(包括其同义词)与所有基准词的相似度都很低，那么可采用 PMI 进行计算。由于已扩展了同义词，因此在计算 PMI 时，就能尽量降低因某些词在语料库中出现频率很低所带来的问题。算法流程如图 1 所示。

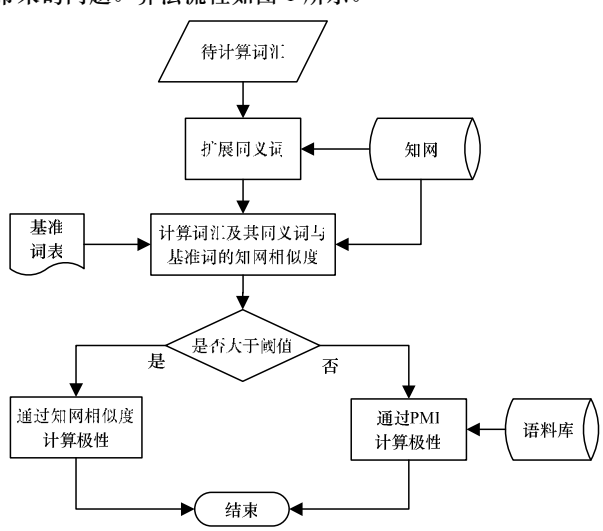

图 1 HowNet 和 PMI 相融合的词语情感极性计算流程

算法描述如下：

(1)待计算极性的词为 $Word_1$，首先通过知网做同义词扩展，假设扩展的同义词集合为 $\{Word_2, Word_3, \cdots, Word_r\}$。

(2)通过式(2)计算 $Word_1$ 及其所有同义词与所有基准词的相似度 $Sim(Word_i,$ 基准词 $_j)$，假设相似度最大的值为 max，预定义一个阈值 θ_{HowNet}，如果 $max>\theta_{\text{HowNet}}$，则表明情感词 $Word_1$(或其同义词)与该基准词的相似度足够高，在大部分情况下可替换使用，因此，$Word_1$ 与该基准词的情感倾向相同，极性 $SO_Hownet(Word_1)$的计算方式为：$SO\_Hownet(Word_1)=Sim(Word_i$,基准词 $_j)*SO($基准词 $_j)$，其中，$SO($基准词 $_j)$表示基准词 j 的极性强度，是预定义的。如果 $Word_1$ 与所有基准词的相似度最大值小于 θ_{HowNet}，表明情感词与所有基准词不可以相互替换，则进行下面步骤的计算。

(3)通过式(1)计算 $Word_1$ 的极性 $SO\_PMI(Word_1)$，预定义一个阈值 θ_{PMI}，如果 $SO\_PMI(Word_1)>\theta_{\text{PMI}}$，则判定为褒义；如果 $SO\_PMI(Word_1)<-\theta_{PMI}$，则判定为贬义。

(4)对于 $Word_1$ 的同义词集合 $\{Word_2, Word_3, \cdots, Word_r\}$，计算所有同义词的极性 $SO\_PMI(Word_2)$, $SO\_PMI(Word_3)$, …, $SO\_PMI(Word_r)$，找出其中$|SO_PMI(Word_i)|$最大的值，预定义一个阈值 θ，如果 $SO\_PMI(Word_i)>\theta$，则判定为褒义；如果 $SO\_PMI(Word_i)<-\theta$，则判定为贬义；否则，判定为中性。

4　实验结果及分析

实验数据来自 COAE(Chinese Opinion Analysis Evaluation)评测库的任务 2。COAE 共设置了 6 个任务[9]，见表 1。

表 1　COAE 任务及说明

任务编号	任务名称	任务说明
1	中文评价词语识别	判断一个词语是否是评价词
2	中文词语的褒贬度分析	判断一个词语的情感倾向
3	中文文本倾向性相关要素的提取	抽取句子中的评价对象
4	中文文本的主客观分类	主客观文本分类
5	中文文本的褒贬度分析	分析文本级的褒贬度
6	中文文本中的观点检索	观点检索

实验首先需要确定一组情感基准词，情感基准词是成对出现的，一对基准词包括一个褒义词和一个贬义词，例如“好”与“坏”，基准词要选取那些褒贬态度明确的、具有代表性的词汇。知网定义的褒义词有 3 866 个，贬义词有 3 261 个，实验的基准词从知网的这些词语中选取，通过计算这些词在语料库中出现的频率，加上人工选取，本文确定 40 对基准词，见表 2。算法中 3 个阈值 θ_{HowNet}、θ_{PMI}、θ 均是根据运行结果观察得到的经验值，本文取值分别为 0.73、0.50、0.20。

表 2　基准词列表

褒义基准词									
美丽	安全	好	正确	温柔	成功	真实	健康	天使	快乐
有益	善良	幸福	廉洁	自信	喜欢	出色	容易	开通	良
便宜	佳	先进	坚强	超常	成熟	完美	有情	最好	著名
卫生	得当	有为	厉害	和平	积极	节俭	优质	高尚	聪明
贬义基准词									
丑陋	危险	坏	错误	暴力	失败	虚假	淫秽	恶魔	伤心
有害	邪恶	悲惨	腐败	自负	讨厌	平凡	困难	封闭	莠
昂贵	差	落后	脆弱	失误	幼稚	缺陷	无情	最差	无名
肮脏	不当	无能	差劲	动荡	消极	浪费	劣质	低俗	愚蠢

采用 F_1 值度量实验结果。实验采用 3 种方法：基于 PMI 的方法，基于知网的方法，本文 PMI 和 HowNet 相融合的方法，实验结果如表 3~表 6，P 表示准确率，R 表示召回率。

表 3　中文词语的褒贬度识别数与正确数分析结果对比

词类	总数	基于 PMI		基于 HowNet		PMI 和 HowNet 融合	
		识别数	正确数	识别数	正确数	识别数	正确数
褒义词	2 678	2 579	2 123	2 819	2 147	2 856	2 349
贬义词	1 919	2 206	1 655	1 782	1 405	1 836	1 554
中性词	1 736	1 548	1 272	1 732	1 429	1 641	1 422

表 4　中文词语的褒贬度 P、R、F_1 分析结果对比

词类	基于 PMI			基于 HowNet			PMI 和 HowNet 融合		
	P	R	F_1	P	R	F_1	P	R	F_1
褒义词	0.823	0.793	0.808	0.762	0.802	0.781	0.822	0.877	0.849
贬义词	0.750	0.862	0.802	0.788	0.732	0.759	0.846	0.810	0.828
中性词	0.822	0.733	0.775	0.825	0.823	0.824	0.867	0.819	0.842

表 5　PMI 和 HowNet 融合方法在的不同阈值下的评测结果

θ_{HowNet}	θ_{PMI}	θ	褒义词			贬义词			中性词		
			P	R	F_1	P	R	F_1	P	R	F_1
0.85	0.75	0.5	0.861	0.719	0.783	0.877	0.708	0.783	0.602	0.890	0.718
0.73	0.75	0.5	0.828	0.741	0.782	0.861	0.726	0.787	0.645	0.865	0.740
0.85	0.5	0.5	0.843	0.744	0.791	0.881	0.759	0.811	0.648	0.865	0.741
0.73	0.5	0.5	0.833	0.807	0.819	0.831	0.750	0.787	0.738	0.851	0.791
0.73	0.5	0.2	0.822	0.877	0.849	0.846	0.810	0.828	0.867	0.819	0.842

表 6　中文词语的褒贬度分析总体评测结果

方法	微平均			宏平均		
	P	R	F_1	P	R	F_1
基于 PMI	0.797	0.797	0.797	0.798	0.796	0.797
基于 HowNet	0.787	0.787	0.787	0.792	0.786	0.789
PMI 和 HowNet 融合	0.841	0.841	0.841	0.845	0.835	0.840

从表 3、表 4 看出，基于 PMI 的词语计算方法对褒义词、贬义词与中性词的表现都较为一般；基于知网的方法对中性词的表现比较高，对褒义词和贬义词效果较差。本文提出的 PMI 和 HowNet 融合方法在褒义词、贬义词和中性词方面比其他 2 种方法有 4%~6%的提升。从表 5 中阈值选取的评测结果可看出，θ_{HowNet}、θ_{PMI}、θ 的变化对结果产生影响：若固定其中的某 2 个阈值，那么当剩余的另一个阈值降低时，就会使得识别为褒、贬义词的词语个数增加，而识别为中性词的词语个数减少，从而导致识别为褒、贬义词的召回率上升、准确率下降，识别为中性词的召回率下降、准确率上升。因此，为了使得识别褒、贬义词和中性词的召回率、准确率和 F_1 值都比较理想，通过实验对比，选取 0.73、0.5、0.2 作

(下转第 193 页)

表 1　不同场景下粒子滤波与本文方法的跟踪结果比较

场景	粒子滤波			本文方法		
	帧速/(f·s^{-1})	*Accuracy*	*Recall*	帧速/(f·s^{-1})	*Accuracy*	*Recall*
高速公路	5.2	0.71	0.72	12.8	0.85	0.91
十字路口	5.1	0.61	0.62	10.3	0.89	0.93
十字路口夜晚	5.4	0.57	0.59	14.3	0.81	0.88

综上所述，本文提出的基于角点动态特征的分层跟踪算法有很多优点，例如可以在灰度图像上进行，对光照变化比较鲁棒，适合夜晚场景，跟踪速度快精度高。但仍然存在不足之处，例如透视效应，图 3(b)中的第 2 幅和图 5(b)中的第 3 幅图像下方的车辆被误认为是 2 个目标。

5　结束语

针对车辆的多目标跟踪中出现的难点，本文结合背景建模思想，设计一种基于角点动态特征的分层实时跟踪方法。利用 Codebook 算法建立背景模型从而提取有效前景，然后在前景区域进行基于角点动态特征的聚类与跟踪。实验结果表明，该方法能对多目标的关联进行较好的匹配，适应复杂多样的交通场景以及多数的天气环境条件，对光照、阴影也有很强的鲁棒性，可应用于交通监控系统。下一步研究方向是解决跟踪性能与算法复杂度之间的平衡问题。

参考文献

[1] Koller D, Weber J, Malik J. Robust Multiple Car Tracking with Occlusion Reasoning[C]//Proc. of the 3rd European Conference on Computer Vision. Stockholm, Sweden: [s. n.], 1994.

[2] Ervin R, MacAdam C, Walker J, et al. System for Assessment of the Vehicle Motion Environment(SAVME)[R]. Washington D. C., USA: National Highway Traffic Safety Administration, Tech. Rep.: UMTRI-2000-21-1, 2000.

[3] Kamijo S, Matsushita Y, Ikeuchi K, et al. Occlusion Robust Tracking Utilizing Spatio-temporal Markov Random Field Model[C]//Proc. of ICPR'00. Barcelona, Spain: IEEE Computer Society, 2000.

[4] Antonini G, Thiran J P. Counting Pedestrians in Video Sequences Using Trajectory Clustering[J]. IEEE Trans. on Circuits and Systems for Video Technology, 2006, 16(8): 1008-1020.

[5] Beymer D, McLauchlan P, Coifman B. A Real Time Computer Vision System for Measuring Traffic Parameters[C]//Proc. of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 1997.

[6] Koller D, Daniilidis K, Nagel H H. Model-based Object Tracking in Monocular Image Sequences of Road Traffic Scenes[J]. International Journal of Computer Vision, 1993, 10(3): 257-281.

[7] Kim Z, Malik J. Fast Vehicle Detection with Probabilistic Feature Grouping and Its Application to Vehicle Tracking[C]//Proc. of IEEE International Conference on Computer Vision. [S. l.]: IEEE Press, 2003.

[8] Worrall A, Sullivan G, Baker K. Pose Refinement of Active Models Using Forces in 3D[C]//Proc. of the 3rd European Conference on Computer Vision. Stockholm, Sweden: [s. n.], 1994.

[9] Stauffer C, Grimson W E L. Adaptive Background Mixture Models for Real-time Tracking[C]//Proc. of IEEE International Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 1999.

[10] 田　峥, 徐　成, 杨志邦, 等. 智能监控系统中的运动目标检测算法[J]. 计算机工程, 2011, 37(4): 1-3.

编辑　金胡考

(上接第 189 页)

为最终的阈值。从表 6 的总体评测结果可知，本文方法在微平均和宏平均方面较其他 2 种方法有约 5%的提升。

5　结束语

传统的基于语料库的词语倾向计算方法依赖于情感词在语料库中的分布规律，对于语料库中出现频率较低的词语的准确率较低；而基于电子词典的计算方法则依赖于电子词典的完善性。针对这 2 种方法的不足，本文提出了一种知网与 PMI 相融合的词语情感极性计算方法，该方法首先利用知网进行同义词扩展来降低情感词在语料库中出现频率低所带来的问题，同时根据知网相似度计算的特性，将知网相似度与 PMI 计算方法相融合。实验结果表明，该方法在微平均和宏平均方面有大约 5%的提升。

本文方法依然有许多改进的余地，例如，文中所使用的基准词是人工选取的，效果尚有提升的余地。下一步将研究更好的基准词选择方法。

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] Wordnet 简介[EB/OL]. [2011-08-23]. http://wordnet.princeton.edu.

[3] Su Fangzhong, Markert K. Subjectivity Recognition on Word Senses via Semi-supervised Mincuts[C]//Proceedings of NAACL'09. Boulder, USA: [s. n.], 2009: 1-9.

[4] 董振东, 董　强. 知网简介[EB/OL]. [2011-08-23]. http://www.keenage.com.

[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, 2005, 20(1): 14-20.

[6] Turney P D, Littman M L. Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus[R]. National Research Council of Canada, Tech. Rep.: EGB-1094, 2002.

[7] 吴泽衡. 基于话题检测和情感分析的互联网热点分析与监控技术研究[D]. 广州: 华南理工大学, 2011.

[8] 姚天昉, 娄德成. 汉语情感词语义倾向判别的研究[C]//第七届中文信息处理国际会议论文集. 北京: 电子工业出版社, 2007: 221-225.

[9] 赵　军, 许洪波, 黄萱菁, 等. 中文倾向性分析评测技术报告[J]. 中国计算机学会通讯, 2008, 4(2): 35-39.

编辑　张正兴