

一种新的中文词语情感极性判别方法

魏 韡^{1,2}, 向 阳²

(1 井冈山大学 电子与信息工程学院, 江西 吉安 343009; 2 同济大学 电子与信息工程学院, 上海 201804)

摘 要: 词语的情感极性判别是文本情感分析的前提和基础. 本文针对如何判别中文词语的情感极性, 提出了一种基于图的半监督学习方法. 该方法将词语看成图中的节点, 链接节点的边的权重由《知网》2008 提供的词语间的语义相似度决定. 少量已知情感极性的词语作为标签, 通过标签传递算法来判别大量未知词语的情感极性. 实验结果表明该方法比传统的基于知网的词语情感极性判别方法有较高的正确率.

关键词: 情感极性; 半监督学习; 图; 知网

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)05-0084-03

A Novel Method to Compute Chinese Word Sentiment Polarity

WEI Wei^{1,2}, XIANG Yang²

(1 College of Electronics and Information Engineering, Jinggangshan University, Ji'an 343009, China;

2 College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Judging the semantic orientation of words is the basic work of the sentiment analysis of text. This paper proposes a graph-based semi-supervised learning to determine the sentiment polarity of Chinese words. The method constructs the graph, whose nodes consist of Chinese words. The weight of two nodes is based on the semantic similarity of two words in HowNet 2008. With the few labeled sentiment words, the polarity of many unlabeled words can be acquired using label propagation algorithm. Experimental results show that proposed method has higher degree of accuracy compared the traditional methods based on HowNet to judging sentiment polarity of words.

Key words: sentiment polarity; semi-supervised learning; graph; HowNet

1 引言

近年来,随着互联网技术尤其是 Web2.0 技术的迅猛发展,越来越多的互联网用户从单纯地获取客观性信息转向主动发布主观性信息. 这些主观性信息蕴藏着丰富的主观情感,由于海量的主观性信息无法靠人工及时处理,借助计算机来自动或半自动地处理和分析主观性信息包含的情感成为当前计算机领域的一个研究热点. 文献[1]对文本情感分析的研究对象和目标进行了定义和说明,并且总结了文本情感分析的研究现状和与进展,同时也指出了

面临的困难和挑战.

文本情感分析涉及到自然语言处理、计算语言学、信息检索、机器学习,人工智能等多个领域,按照分析的粒度不同,可大致分成词语级情感分析、句子级情感分析、篇章级情感分析三个研究层次. 其中词语级情感分析是整个文本情感分析的基础.

2 相关的研究工作

词语级情感分析的主要任务是判别词语的褒贬性或称为极性. 具有情感极性的词语大部分是形容词或副词,也包括一部分的名词或动词. 为了定量地

收稿日期: 2012-07-28; **修回日期:** 2012-08-22

基金项目: 国家自然科学基金项目(61103069,71171148);江西省教育厅青年科学基金项目(GJJ11178);吉安市软科学计划项目(吉市科计字[2012]32-3)

表示词语的极性强度,通常用位于区间 $[-1,1]$ 之间的某个实数作为情感权重表示词语的极性强度.情感权重大于0,则词语为褒义词;情感权重小于0,则词语为贬义词.情感权重的绝对值越大则意味着词语的极性强度越大.词语的情感极性判别主要有基于语料库的统计方法和基于词典的语义相似度两种方法.

2.1 基于语料库的统计方法

基于语料库统计的方法主要利用在大规模语料库中挖掘出的语言学规则或通过机器学习获得的语言模型来对词语的情感极性进行判别^[2-4].基于语料库统计的方法依赖于语料库的规模和质量,而且计算比较复杂,效率不高.

2.2 基于词典的语义相似度方法

基于词典的语义相似度方法是利用英文词典 WordNet 或中文词典 HowNet(知网)提供的语义相似度来判别词语的情感极性^[5-6].该方法有较高的准确率和较好的扩展性.基于词典的语义相似度方法依赖于构建成熟的词典,所以准确度比较高,但对种子词的选择会影响到计算的结果.

3 基于图的半监督学习方法

我们借鉴文献[7]基于图的半监督学习方法,利用少量已知情感极性的词语来推断大量待判别词语的情感极性.

3.1 概念定义及算法流程

定义1 无向图.由全部词语构成的无向图记为 $G=(V, E)$,其中 V 表示图中词语节点的集合, E 表示链接图中词语的边的权重集合.

定义2 边的权重.边的权重 w_{ij} 由边链接的两个词语节点 i 和 j 之间基于 HowNet 的语义相似度决定.

定义3 词语的标签.由于词语的情感极性可分为褒义和贬义,所以词语的标签分为两类,即褒义词的标签为+1,贬义词的标签为-1.

定义4 已知标签节点类别矩阵.已知标签的节点 x_i 由 1×2 的行向量 f_i 表示.如果 x_i 的标签是+1,则 f_i 为 $[1, 0]$;如果 x_i 的标签是-1,则 f_i 为 $[0, 1]$.已知标签节点类别矩阵由已知标签的节点的行向量组成,记为 f_L .

定义5 未知标签节点类别矩阵.未知标签的节点 x_k 由 1×2 的行向量 f_k 表示,其中 f_{kj} ($j=1$ 或 2)可随机取 $[0, 1]$ 之间任何一个实数.未知标签节点类别矩阵由未知标签的节点的行向量组成,记为 f_U .

定义6 所有节点类别矩阵.所有节点类别矩阵由已知标签节点类别矩阵和未知标签节点类别矩阵组成,记为 f_X ,即 $f_X = \begin{bmatrix} f_L \\ f_U \end{bmatrix}$.

基于图的半监督学习方法基于流行假设,即局部领域内的数据具有相似的特性,其标记也相似.这个假设和以下假设是相吻合的,即词语之间语义相似度越大,词语的情感极性越一致.首先为每个节点定义一个类别的概率分布,表示该节点的类别标记.每次传播的过程中,每个节点都通过其相邻节点的概率分布来更新自己的概率分布.由于与某节点相似度越大的相邻节点对其的概率分布的影响越大,这样相似的节点的概率分布也将趋于相似直到收敛到平衡状态.在传播的过程中要保持已知节点的类别标记不变,即保持其概率分布不变.具体的算法流程如下:

(1)根据定义1和定义2构建图 G ,其中权重矩阵 $W = [w_{ij}]$.

(2)根据定义4~定义6,初始化 f_X ,其中 $\sum_{j=1}^2 f_{kj} = 1$.

(3)计算转移概率矩阵 $P = [p_{ij}]$,其中 p_{ij} 表示节点 p_i 从节点 p_j 获得标记的概率,定义为:

$$p_{ij} = p(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^n w_{kj}} \quad (1)$$

(4)计算新的 f_X ,即 $f_X \leftarrow P \times f_X$

(5)对 f_U 进行归一化处理,即保证 $\sum_{j=1}^2 f_{kj} = 1$.

(6)将 f_L 修改为初始值 F_L .

(7)重复步骤(4)~(6)直到 f_U 收敛.

(8)未知标签节点 x_k 的类别由 $\arg\max f_{kj}$ 决定.

在算法流程的(3)至(7)步骤中,未知词语的类别标记被不断地更新,而已知词语的类别标记则保持不变.步骤(6)保证已标记的节点在传递的过程中被强化,维持标记的“强度”而不被“淡化”.类别标记以标签节点为起点,最先传递到其近邻的未标签节点,然后再传递给次近邻的未标签节点.整个过程最终将收敛到平衡状态,即所有节点的标记不再改变,从而得到未标签节点的分类标记.

3.2 收敛性证明

本小节将证明 f_X 在传播过程中将会收敛.首先将转移概率矩阵 P 依据有标签节点的个数和未标签节点的个数表示如下: $P = \begin{bmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{bmatrix}$

传播过程的关键是

$$f_U \leftarrow P_{UL} f_L + P_{UU} f_U \quad (2)$$

若 f_U 最终将收敛, f_U^0 表示成 f_U 的初始值, 公式(2)可以表示成

$$f_U = \lim_{n \rightarrow \infty} (P_{UU})^n f_U^0 + \left(\sum_{i=1}^n (P_{UU})^{(i-1)} \right) P_{UL} F_L \quad (3)$$

因为 P 是行归一化矩阵, P_{uu} 是 P 的子矩阵, 故

$$\exists \beta < 1, \sum_{j=1}^u (P_{UU})_{ij} \leq \beta, \forall i = 1 \dots u, \text{可以得到}$$

$$\begin{aligned} \sum_j (P_{UU})_{ij}^n &= \sum_j \sum_k (P_{UU})_{ik}^{(n-1)} (P_{UU})_{kj} \\ &= \sum_k (P_{UU})_{ik}^{(n-1)} \sum_j (P_{UU})_{kj} \\ &\leq \sum_k (P_{UU})_{ik}^{(n-1)} \beta \\ &\leq \beta^n \end{aligned}$$

因此, 矩阵 P 的行和收敛至 0, 即 $(P_{UU})^n f_U^0 \rightarrow 0$, 同时也证明了 f_U 的初始值不影响收敛结果. 最后公式(3)可表示成:

$$f_U = (I - P_{UU})^{-1} P_{UL} F_L \quad (4)$$

由此可说明该算法将收敛于一个确定值.

4 实验与分析

本文选取文献[5]中 40 对褒贬基准词作为已知情感极性的词语, 采用 HowNet2008 提供的正面情感词和负面情感词作为测试集 1, 正面评价词和负面评价词作为测试集 2, 测试集 1 和测试集 2 所有词语组成测试集 3. 表 1 给出了各测试集的信息.

表 1 测试集

测试集编号	褒义词个数	贬义词个数
1	873	1297
2	3866	3261
3	4739	4558

本文将已知情感极性的词语和测试集中的词语构成无向图, 节点之间的权重借助 HowNet2008 提供的词语相似度计算的软件包来计算得到. 实验设定褒义基准词的标签是 +1, 贬义基准词的标签是 -1. 通过基于图的半监督算法得到未标签词语的标记, 标签是 +1 的词语判别是褒义词, 标签是 -1 的词语判别是贬义词. 本文用判别正确率来验证算法的有效性. 判别正确率的计算公式如下:

$$accuracy = \frac{number(wr_i)}{number(wt_i)} \quad (5)$$

其中 $number(wr)$ 表示情感极性判别正确的词

语个数, $number(wt)$ 由测试集的词语个数扣除基准词的个数得到, i 表示测试集的编号. 使用 40 对基准词, 分别采用了文献[7]的语义相似度方法和基于图的半监督学习方法在 3 个测试集上实验, 实验结果如表 2 所示.

表 2 两种方法的判别准确率

测试集编号	文献[7]的方法	本文的方法
1	75.2%	91.6%
2	72.6%	88.5%
3	73.2%	90.1%

实验结果表明, 基于图的半监督学习的方法比文献[7]的方法能够较大幅度地提高判别的准确率, 证明了本文方法的有效性. 为了观察不同的基准词个数对判别准确率的影响, 本文将 40 对基准词语在 Google 搜索中返回的页面数作为该词语的频率, 选取频率最高的 10 对作为第一组, 频率最高的 20 对作为第二组, 频率最高的 30 对作为第三组, 所有的 40 对作为第四组, 分别测试了四组基准词在 3 个测试集对结果的影响, 实验结果如图 1 所示.

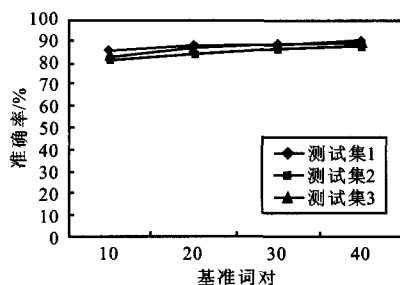


图 1 基准词的数量对实验的影响

从图 1 可以看出, 随着基准词个数的增加, 准确率都有相应的提高, 表明基准词个数对准确率有影响. 但是当基准词的个数达到 30 对时, 准确率就比较稳定了, 表明基于图的半监督学习方法对基准词的个数不是很敏感.

5 结束语

本文提出了一种基于图的半监督学习方法来判别中文词语的情感极性. 该方法利用 HowNet2008 提供的词语间的语义相似度来构建词语的无向图, 使用少量已知情感极性的词语作为标签来判别大量未知标签词语的情感极性. 实验结果表明了该方法的有效性. 今后的工作, 可考虑选取更具代表性和情感极性更强烈的词语作为标签, 同时考虑词语之间其他关系作为链接的权重.

(下转第 91 页)

延时 T 大小对电路性能的影响. 通过对传输函数基于拉普拉斯变换的理论分析以及 Matlab 仿真, 分析电路延时参数存在一个最优化值. 基于 SPICE 对电路与互连线模型进行仿真, 分析输出信号眼图与误码率, 证明本文的优化设计方法可以有效地改善电路性能, 提高了信道传输带宽, 抑制码间干扰, 实现片上互连的高速传输.

参考文献:

- [1] S Vangal, J Howard, G Ruhl, et al. An 80—tile 1.28TFLOPS network—on—chip in 65 nm CMOS[C]//IEEE ISSCC Dig Tech. San Francisco: IEEE, 2007: 98—99.
- [2] Daniel Schinkel. Low—power, high—speed transceivers for Network—on—chip communication[J]. Very Large Scale Integration (VLSI) Systems, IEEE Transactions, 2009, 17(1): 12—21.
- [3] Bai Yun. High—speed energy—efficient on—chip interconnect driver and receiver[D]. US: Stanford University, 2008.
- [4] Ron Ho, Tarik Ono, Robert David Hopkins. High speed and low energy capacitively driven on—chip wires[J]. IEEE Journal of Solid—State Circuit, 2008, 43(1): 52—60.
- [5] Richard T Chang, Niranjana Talwalkar. Near speed—of—light signaling over on—chip electrical interconnects[J]. IEEE Journal of Solid—State Circuit, 2003, 38(5): 834—838.
- [6] Daniel Schinkel, Eisse Mensink, Eric A M Klumperink. A 3—Gb/s/ch Transceiver for 10—mm Uninterrupted RC—Limited Global On—Chip Interconnects[J]. IEEE Journal of Solid—State Circuit, 2006, 41(1): 297—306.
- [7] Liang Zhang, John M Wilson, Rizwan Bashirullah, et al. Voltage—mode driver preemphasis technique for on—chip global Buses[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2009, 15(2): 231—236.
- [8] R Bashirullah, L Wentai, R Cavin. III, Delay and power model for currentmode signaling in deep submicron global interconnects[C]// Custom Integrated Circuits Conference Proc of the IEEE. [s. l.]: IEEE, 2002: 513—516.
- [9] James F Buckwalter, Mounir Meghelli, Daniel J Friedman, et al. Phase and amplitude pre—emphasis techniques for low—power serial links[J]. IEEE Journal of Solid—State Circuits, 2006, 41(6): 1391—1399.
- [10] Shi Xinzhi, Qi Chang. A pade approximation approach for generalized inverse matrix exponential function[J]. Journal of Huazhong University of Science and Technology (Nature Science Edition), 2008(8): 49—55.
- [11] Ribeiro N S, Conforti E, Gallep CM. BER estimation from measured eye diagrams of a wavelength regenerative converter employing a single—SOA[C]// SBMO/IEEE MIT—S International Microwave and Optoelectronics Conference (IMOC). Brazil: IEEE, 2009: 655—659.

作者简介:

楼媛女, (1987—), 硕士. 研究方向为视频编解码算法的研究和硬件实现、多媒体加速器架构设计.

(上接第 86 页)

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] WIEBE J. Learning subjective adjectives from corpora [C]//Proceedings of the 17th National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2000: 735—740.
- [3] Yuen R W M, Chan T Y W, Lai B Y, et al. Morpheme—based derivation of bipolar semantic orientation of Chinese words[C] //Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: IEEE, 2004: 1008—1014.
- [4] Turney, Littman. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315—346.
- [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): 14—20.
- [6] 李纯, 乔保军, 曹元大, 等. 基于语义分析的词汇倾向识别研究[J]. 模式识别与人工智能, 2008, 21(4): 482—487.
- [7] Zhu X J, Ghahramani. Learning from labeled and unlabeled data with label propagation[R]. Carnegie Mellon University, 2002.

作者简介:

魏韡 (1983—), 男, 博士研究生, 讲师. 研究方向为文本情感分析、人工智能.

向阳 (1962—), 男, 教授, 博士生导师. 研究方向为数据挖掘、人工智能.