文章编号:1007-5321(2010)05-0141-04

PageRank 模型在中文情感词极性判别中的应用

李荣军, 王小捷, 周延泉
(北京邮电大学 计算机学院, 北京 100876)

摘要: 针对倾向性分析任务中的基础性工作——情感词的极性判断工作,提出了一种基于 PageRank 模型的情感词极性判断方法. 由待判别情感词和少量种子情感词构成图中的结点,利用知网(HowNet)语义资源计算词语间的语义相似度,进而得到图中结点间边的权重. 通过 PageRank 模型的引入,综合利用有标种子情感词和无标待判别情感词实现对无标情感词的极性判别. 与传统的基于 HowNet 的情感词判别方法相比,PageRank 模型的引入使情感词判别的准确率平均提高 10% 左右,充分验证了所提方法的可行性.

关 键 词: 自然语言处理; 语义倾向分析; PageRank 模型; 知网
中图分类号: TP391 **文献标志码**: A

Semantic Orientation Computing Using PageRank Model

LI Rong-jun, WANG Xiao-jie, ZHOU Yan-quan
(School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: For determining the polarity of sentiment words, an algorithm based on PageRank technology is proposed. A graph is constructed whose nodes consist of unlabeled sentiment words and a few sentiment seeds, and the weights between each two nodes based on the semantic similarity of HowNet are also gained. With the PageRank technology on those seeds the polarity of the unlabeled sentiment words can be obtained. Compared with the methods based on HowNet to judging polarity of sentiment words, the proposed algorithm of combining PageRank technology shows its effectiveness by 10% increase of the precision.

Key words: natural language processing; semantic orientation; PageRank model; HowNet

0 引言

文本的情感倾向性分析在人机交互、问答系统、舆论监督、人性检索等领域有广泛应用,已成为当前自然语言处理领域研究的热点问题. 词汇的倾向性(极性)识别任务作为倾向性分析系统中的基础性工作,更得到了极大关注.

情感词极性判别方法可分为 2 类:基于大规模语料库的统计方法和利用人工构建的语义知识库计算相似度方法.

基于统计的方法主要利用在大规模语料中挖掘出的语言学规则或习得的机器学习模型对词汇的情感极性进行判别. 例如,文献[1-2]以情感词间的连接关系作为特征来推断情感词在某领域内的极性;文献[3]依靠在特殊关系句型中的词同现规则,对评价词的情感极性作出判断;文献[4]预先挑选若干具有较明确极性信息的形容词构成种子词集合,通过计算待测词与褒、贬种子极性词之间的点态互信息差值(SO-PMI)来确定待测词汇的极性;语义知识库的建立,给极性分析工作的开展以极大支持.

收稿日期: 2010-02-05
基金项目: 国家自然科学基金项目(90920006); 国家教育部博士点基金项目(20090005110005)
作者简介: 李荣军(1982—), 男, 博士生, E-mail: lirongjun2002@bupt.edu.cn; 王小捷(1969—), 男, 教授, 博士生导师.

目前,最典型的语义知识库有多语言知识库 WordNet 和汉语言知识库 HowNet(《知网》). 文献[5]尝试利用 WordNet 构建词汇相似度来分析形容词的极性;在此基础上,文献[6]利用 WordNet 中的词汇关系和词汇解释信息帮助判断词汇极性;文献[7]将情感词极性判别问题等价于在图中进行标号传递的半监督问题. 通过引入最小割和标号传播算法结合 WordNet 和 OpenOffice 分别对英语和法语的情感词极性进行分析;文献[8]提出基于 HowNet 词汇语义相似度和语义相关场的情感词极性计算方法,其中基于语义相似度的方法收到了较理想的效果.

本文借鉴于 PageRank 模型[9]和跨领域倾向性分析算法[10],通过引入 PageRank 模型来研究词汇的原始情感倾向性识别技术,即判断在通常情况下,情感词表现出的情感极性. 利用 HowNet 资源得到情感词之间的连接权重,通过情感种子词与待测情感词、待测情感词与待测情感词之间的语义相似度关系解决待测情感词的极性判别问题. 另外,本文通过理论证明文献[10]针对测试集所采用的"伪标签"的值,在实际应用中可任意取定.

1 基于 PageRank 模型的情感词极性判别

1.1 问题描述

种子情感词集合向量为 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_n\}$,其手工标注的情感极性向量为 $\boldsymbol{Y}_S=\{y_1,y_2,\cdots,y_n\}$;待分类情感词向量为 $\boldsymbol{W}=\{w_{n+1},w_{n+2},\cdots,w_{n+m}\}$,待求标注结果为 $\boldsymbol{Y}_W=\{y_{n+1},y_{n+2},\cdots,y_{n+m}\}$. 当情感词属于褒义情感词时,$y_i=1$;反之,$y_i=-1$. 一方面,必须依赖种子情感词提供的极性信息 $\boldsymbol{Y}_S$ 对待分类情感词的情感极性进行判断;另一方面,考虑到具有相同极性的情感词通常具有强烈的语义相似性,反之,具有相反极性的情感词之间的语义相似性较弱,因此,期待待分类情感词之间的语义相似关系对于判断其极性也会有一定的帮助.

1.2 算法描述

定义图 $G=<\boldsymbol{N},\boldsymbol{M}>$,$|\boldsymbol{N}|=|\boldsymbol{S}|+|\boldsymbol{W}|$,其中 $\boldsymbol{N}$ 为 G 中的结点集合(结点由全部情感词构成),$|\boldsymbol{S}|$ 为种子情感词数,$|\boldsymbol{W}|$ 为待分类情感词数,$|\boldsymbol{W}|\times|\boldsymbol{N}|$ 连接矩阵 $\boldsymbol{M}$ 描述结点间的无相图链接关系,M_{ij} 为结点 i 与结点 j 间的语义相似度,$\boldsymbol{M}$ 可分解为 $|\boldsymbol{W}|\times|\boldsymbol{S}|$ 的子矩阵 $\boldsymbol{U}$ 和 $|\boldsymbol{W}|\times|\boldsymbol{W}|$ 的子矩阵 $\boldsymbol{V}$ 共2部分,U_{ij} 为待测情感词 i 和种子情感词 j 之间的语义相似度,V_{ij} 为待测情感词 i 和待测情感词 j 之间的语义相似度. 引入 PageRank 模型后,情感词极性判别算法的迭代公式如下

$$\boldsymbol{Y}_W^{(n)}=(1-\beta)\boldsymbol{U}\,\boldsymbol{Y}_S+\beta\boldsymbol{V}\,\boldsymbol{Y}_W^{(n-1)} \tag{1}$$

其中,$\boldsymbol{Y}_W^{(n)}$ 为第 n 次迭代后的 $\boldsymbol{Y}_W$,β 为加权系数,$0<\beta<1$.

考虑到一些来自与待测情感词连接权重较低的结点的"投票"可能是不可靠的,同时为保证迭代过程收敛,需对 $\boldsymbol{U}$、$\boldsymbol{V}$ 和 $\boldsymbol{Y}$ 作相应处理,算法流程描述如下.

1) 构建 $\boldsymbol{M}$. 其中,连接权重利用 HowNet 语义相似度计算给出[7].

2) 计算 $\max k(V_{i*})$,$1<i<|\boldsymbol{W}|$. $\max k(V_{i*})$ 为 $\boldsymbol{V}$ 中第 i 行各元素按值由大到小排序,排名在第 k 位元素的值.

3) $\forall i,j,\exists V_{i,j}\leqslant\max k(V_{i*})$ 时,$V_{i,j}=0$;否则,$V_{i,j}$ 保持不变.

4) 对 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 进行联合归一化,

$$\left.\begin{aligned}U_{ij}&=\frac{U_{ij}}{\sum_{j=1}^{|\boldsymbol{S}|}U_{ij}+\sum_{j=1}^{|\boldsymbol{W}|}V_{ij}}\\ V_{ij}&=\frac{V_{ij}}{\sum_{j=1}^{|\boldsymbol{S}|}U_{ij}+\sum_{j=1}^{|\boldsymbol{W}|}V_{ij}}\end{aligned}\right\} \tag{2}$$

5) 对 $\boldsymbol{Y}_S$ 进行归一化,以保证 $\boldsymbol{Y}_S$ 中褒义情感词分值和为 +1,贬义情感词分值和为 −1.

$$y_i=\begin{cases}-\dfrac{y_i}{\sum_{i=1}^{|\boldsymbol{S}|}y_i}, & y_i<0\\ \dfrac{y_i}{\sum_{i=1}^{|\boldsymbol{S}|}y_i}, & y_i>0\end{cases}\qquad i=1,2,\cdots,|\boldsymbol{S}| \tag{3}$$

6) 设置迭代初始值 $\boldsymbol{Y}_W^{(0)}$.

7) 利用式(1)计算 $\boldsymbol{Y}_W$.

8) 对 $\boldsymbol{Y}_W$ 进行如下归一化

$$y_i=\begin{cases}-\dfrac{y_i}{\sum_{i=|\boldsymbol{S}|+1}^{|\boldsymbol{S}|+|\boldsymbol{W}|}y_i}, & y_i<0\\ \dfrac{y_i}{\sum_{i=|\boldsymbol{S}|+1}^{|\boldsymbol{S}|+|\boldsymbol{W}|}y_i}, & y_i>0\end{cases}$$

$$i=|\boldsymbol{S}|+1,|\boldsymbol{S}|+2,\cdots,|\boldsymbol{S}|+|\boldsymbol{W}| \tag{4}$$

保证 $\boldsymbol{Y}_W$ 中褒义的情感词分值和为 +1,贬义的情感词分值和为 −1. 转步骤7),直到 $\boldsymbol{Y}_W$ 的值收敛为止.

9）依据 $\boldsymbol{Y}_W$的值对情感词 w_i的极性作出判断.

$$w_i=\begin{cases}\text{褒义}, & y_i>T\\ \text{贬义}, & y_i<T\end{cases}$$

$$i=|\boldsymbol{S}|+1,|\boldsymbol{S}|+2,\cdots,|\boldsymbol{S}|+|\boldsymbol{W}| \qquad (5)$$

y_i的绝对值大小表示词 i 褒贬强烈程度.

算法对于步骤 6）中涉及的迭代初始值 $\boldsymbol{Y}_W^{(0)}$ 的选取是不敏感的，证明过程如下.

$$\boldsymbol{Y}_W^{(n)}=(1-\beta)\boldsymbol{U}\boldsymbol{Y}_S+\beta\boldsymbol{V}\boldsymbol{Y}_W^{(n-1)}$$

令$(1-\beta)\boldsymbol{U}\boldsymbol{Y}_S=\boldsymbol{\varphi}$，则

$$\boldsymbol{Y}_W^{(n)}=(\beta\boldsymbol{V})^n\boldsymbol{Y}_W^{(0)}+\sum_{i=1}^{n}(\beta\boldsymbol{V})^{i-1}\boldsymbol{\varphi}$$

$$\lim_{n\to\infty}\boldsymbol{Y}_W^{(n)}=\lim_{n\to\infty}\left[(\beta\boldsymbol{V})^n\boldsymbol{Y}_W^{(0)}+\sum_{i=1}^{n}(\beta\boldsymbol{V})^{i-1}\boldsymbol{\varphi}\right]$$

由 $\boldsymbol{V}$ 的定义可知，$\boldsymbol{V}$ 中所有元素都大于 0，此外，由于 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 进行过联合行规范化，并且

$$\sum_j(\beta\boldsymbol{V})_{ij}^n=\sum_j\sum_k(\beta\boldsymbol{V})_{ik}^{n-1}(\beta\boldsymbol{V})_{kj}=\sum_k(\beta\boldsymbol{V})_{ik}^{n-1}\sum_j(\beta\boldsymbol{V})_{kj}<\beta\sum_k(\beta\boldsymbol{V})_{ik}^{n-1}$$

所以

$$\sum_j(\beta\boldsymbol{V})_{ij}^n<\beta^n,\ \forall i=1,2,\cdots,|\boldsymbol{W}|$$

$$\lim_{n\to\infty}=(\beta\boldsymbol{V})^n\boldsymbol{Y}_W^{(0)}=0$$

因此，序列 $\boldsymbol{Y}_W^{(n)}$ 是收敛的. 即 $\boldsymbol{Y}_W^{(0)}$ 的选取不会影响对 $\boldsymbol{Y}_W^{(n)}$ 的估计.

2　实验与分析

2.1　实验设置

选取 HowNet 情感词典中的 814 条正面情感词和 1 232 条负面情感词作为实验语料. 为了从结果曲线中直观地观察到算法分别对褒、贬义词条的判别性能，在对实验结果评测时，将褒义情感词排在前面，贬义情感词排在后面. 将 $\boldsymbol{Y}_W^{(0)}$ 初始化为零向量，采用 0 作为默认阈值. 种子词全集选用文献[8]所收集的 40 组褒贬基准词(见表 1). 共选取 4 组种子词子集以便于分析算法对于种子词数变化的敏感程度. 实验中，同时考察 β 和 k 的取值对算法性能的影响.

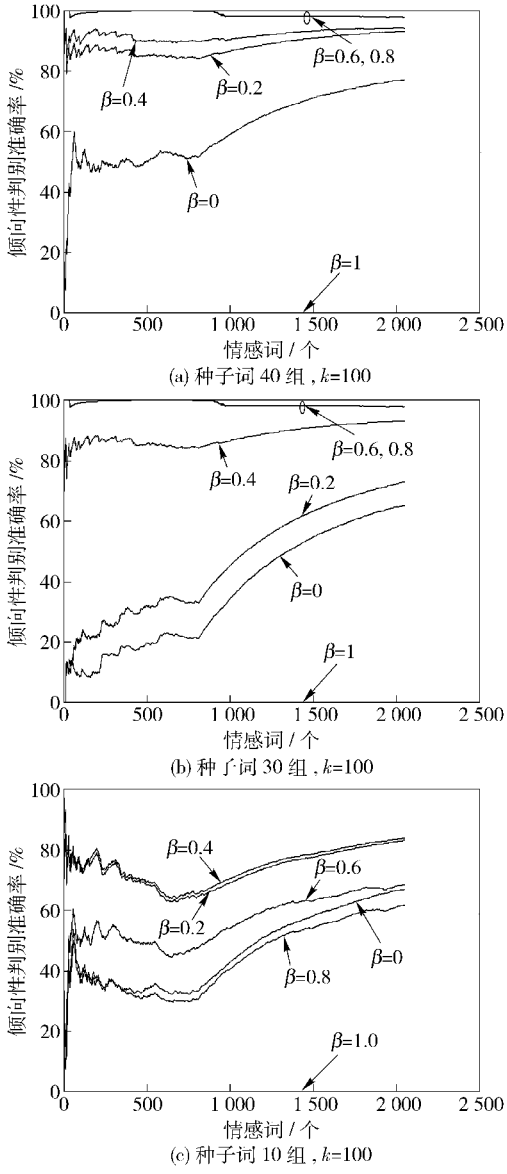
表 1　40 组褒贬基准词

褒义基准词									
健康	安全	天下第一	美丽	超级	保险	卫生	天使	英雄	精选
快乐	权威	稳定	优秀	高级	精英	最好	最佳	幸福	容易
高手	文明	积极	著名	漂亮	完美	简单	和平	开通	真实
先进	便宜	优质	欢乐	美好	良好	不错	出色	成熟	完善
贬义基准词									
不合作	黑客	疯狂	错误	事故	非法	失败	背后	麻烦	不良
病人	恶意	色情	暴力	黄色	浪费	落后	漏洞	有害	讨厌
自负	不安	魔鬼	花样	野蛮	陷阱	不当	和平	开通	真实
淫秽	流氓	虚假	残酷	变态	脆弱	不合格	愚人	恶劣	恶魔

2.2　参数敏感性分析

2.2.1　β 敏感性分析

当 $\beta=0$ 时，表明算法中不考虑无标待测情感词之间的语义相似度对其自身极性判断过程的影响，此时算法退化为文献[8]算法，将此时的实验曲线作为基线. 图 1 示出了 β 的敏感性分析曲线图. 由图 1 可见，基线的判别方法对种子词数和质量的选择较敏感. 当种子词选取数逐渐减少时，基线方法的判别准确率下降趋势较明显. 另外，基线方法对



(a) 种子词 40 组，k=100

(b) 种子词 30 组，k=100

(c) 种子词 10 组，k=100

图 1　β 敏感性分析

褒、贬情感词语判别的准确率也存在较大差别,这很大程度上同样是由于种子词的选取质量造成的. 尽管可以保证选取的褒、贬义种子词在数量上对等,但很难得知褒、贬义种子词的质量是否得到了保障. 可以看到,当β取值适当增大时,词语极性的判别准确率有较大上升,有时甚至能达到95%以上,同时对于褒、贬情感词语的判别准确程度也显得比较平均. 当然,β不能无限增大,当$\beta=1$时,由于没有引入标注的样本作为指导,倾向性判断的结果为$\boldsymbol{Y}_W$的初始值零向量. 通常的情况下取$\beta<0.5$为宜,即强调种子情感词的指导作用.

2.2.2 k敏感性分析

图2示出了k敏感性分析曲线图. 由图2可以看出,k的取值对算法有效性产生较大影响. 当k值较大时,即在判别当前的情感词时,如果过度考虑一些与它连接权重较低的情感词对其极性判别的影响,会导致极性判别的准确率大幅度下降. 这可能由于HowNet对于词语相似度的辨别精度是有限的,如果在算法中引入了过多较小的相似度权值,就等价于对噪声进行了累积.

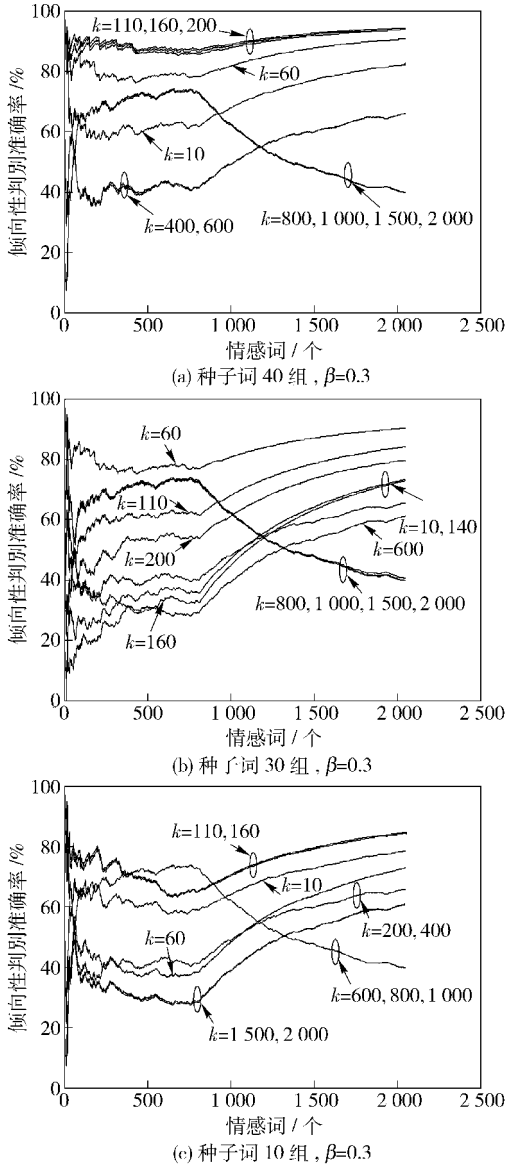


图2 参数K敏感性分析

3 结束语

本文提出了一种基于PageRank模型的情感词极性判断方法. 区别于单纯的利用有标语料对待测情感词进行判别的方法,PageRank模型的引入,使无标情感词集对其自身的极性判别过程也产生了一定的影响. 实验结果充分验证了本文方法的可行性. 需要强调,虽然在算法描述中对图结点的连接权重是利用HowNet计算得到的,但是该连接权重也可通过其他有监督或无监督的方法得到.

参考文献:

[1] Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives[C]// ACL 1997. Stroudsburg: [s. n.], 1997: 174-181.

[2] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[C]// EMNLP 2006. Sydney: [s. n.], 2006: 355-363.

[3] Popescu A, Etzioni O. Extracting product features and opinions from reviews[C]// HLT-EMNLP 2005. Vancouver: [s. n.], 2005: 339-346.

[4] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]// ACL 2002. Philadelphia: [s. n.], 2002: 417-424.

[5] Hu Minqing, Liu Bing. Mining and summarizing customer reviews[C]// ACM SIGKDD. Seattle: [s. n.], 2004: 168-177.

[6] Andreevskaia A, Bergler S. Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses[C]// EACL 2006. Trento: [s. n.], 2006: 209-216.

[7] Rao D, Ravichandran D. Semi-supervised polarity lexicon induction[C]// EACL 2009. Athens: [s. n.], 2009: 675-682.

[8] 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.
Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[9] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web[R]. Stanford: Stanford University, 1999.

[10] 吴琼, 谭松波, 张刚, 等. 基于图排序模型的跨领域倾向性分析算法[C]// 中国计算机语言学研究前沿进展. 烟台: [s. n.], 2009: 618-623.