

一种新闻评论情感词典的构建方法

周咏梅¹ 阳爱民¹ 杨佳能²

(广东外语外贸大学思科信息学院 广州 510006)¹

(广东外语外贸大学国际工商管理学院 广州 510006)²

摘 要 情感词典研究是文本情感分析领域的一个重要内容;基于情感词典的文本情感分析方法是一种非常有效的方法。互联网上的新闻评论包含评论人的情感,对其情感进行自动分析研究是非常有意义的。借鉴图排序模型的原理,提出一种新闻评论情感词典构建方法,该方法首先通过新闻评论语料和基础情感词典获得评论情感词集和种子词,然后根据提出的基于 PageRank 算法的方法判定评论情感词集的极性并计算其强度,进而构建新闻评论情感词典。实验从情感词判定的准确性和基于构建的情感词典的分类性能两个方面验证了所提方法的有效性。

关键词 情感词典,图排序,PageRank 算法,新闻评论

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.014

Construction Method of Sentiment Lexicon for News Reviews

ZHOU Yong-mei¹ YANG Ai-min¹ YANG Jia-neng²

(Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, China)¹

(School of Management, Guangdong University of Foreign Studies, Guangzhou 510006, China)²

Abstract Research on sentiment lexicon is an important content on text sentiment analysis research field. The method of text sentiment analysis based on sentiment lexicon is very effective. It is very meaningful on automatic analysis sentiment for news reviews on internet. In this paper, a construction method of sentiment lexicon for news reviews was proposed, which learns Graph-Ranking model principle. The basic idea of the proposed method is as follows. Firstly, Get the word set of news reviews and the seed words with the corpus and basic sentiment lexicon. Then calculate the strength and polarity of the word set of news reviews according to the proposed algorithm based on PageRank. Finally, Construct the sentiment lexicon for news reviews. From two aspect the experiments show that the proposed method is effective, one is the accuracy of deciding the polarity of sentiment words, the other is the classification performance using the sentiment lexicon which was constructed by proposed method.

Keywords Sentiment lexicon, Graph-ranking, PageRank algorithm, News reviews

1 引言

随着互联网的快速发展,人们越来越习惯于在网上表达自己对日常事件、新闻、产品、政策等的观点和看法,这使得网上存在大量带有情感(Sentiment)倾向性的文本。如何对浩如烟海的网络文本进行快速的情感倾向性分析,就成为越来越引起广泛关注的研究问题。

在情感倾向分析的相关研究中,情感词典(Sentiment Lexicon)发挥了重要的作用。利用高质量的情感词典,实际的应用系统采用简单快速的方法就可以得到很好的效果^[1-4]。但是,网络语言呈现出时效性、领域性和不规范的特点,手工构建词典不仅费时费力,而且需要动态更新。因此,许多研究致力于采用自动的方法进行情感词典的构建,包括使用词语资源^[5,6]、大型语料库^[7]甚至互联网^[8]来获取词语之间的语

义关系,从而构建出大规模的情感词典。情感词典按照构建方式的不同又可以分为弱监督学习、半监督学习和监督学习3种方式。文献[7]从大语料库华尔街日报中发掘出大量的形容词性的评价词语,应用聚类的方法将形容词划分为褒义和贬义两类;Turney等^[9]使用了点互信息(Pointwise Mutual Information, PMI)和潜在语义分析(Latent Semantic Analysis, LSA)两种统计方法,通过搜索引擎和语料库计算词语和种子词语之间的点互信息,从而得到词语的极性。

本文以中文新闻评论情感分析为研究对象,借鉴图排序算法(如 PageRank^[10])的思想(图中与其它重要结点紧密相连的结点也很重要),将词的情感与图排序算法结合起来,提出基于 PageRank 算法的中文新闻评论情感词典构建方法。PageRank 算法执行随机冲浪模型,收敛速度快,已成功应用于很多领域。本文提出的这种情感词典构建方法首先通过新

到稿日期:2013-06-01 返修日期:2013-07-25 本文受国家社科基金项目(12BYY045),教育部人文社会科学研究青年项目(10YJCZH247),广东省科技计划项目(2010B031000014),教育部新世纪优秀人才支持计划(NCET-12-0939)资助。

周咏梅(1971—),女,教授,主要研究方向为文本情感分析、微博情感分析, E-mail: yongmeizhou@gdufs.edu.cn; 阳爱民(1970—),男,博士,教授,主要研究方向为模式分类、文本情感分析、机器学习; 杨佳能(1987—),男,硕士生,主要研究方向为文本情感分析、机器学习。

闻评论语料和基础情感词典获得评论情感词集与种子词,并使用 PMI 算法计算评论情感词集与种子词的相似矩阵,然后根据提出的基于 PageRank 算法的方法判定评论情感词集的极性并计算其强度,进而构建新闻评论情感词典。

本文第 2 节主要介绍提出的情感词典构建方法的基本思路;详细的词典构建方法在第 3 节描述;第 4 节介绍实验,并分析提出方法的有效性;最后总结全文。

2 新闻评论情感词典构建思路概述

本文提出的中文新闻评论情感词典构建的基本思路如图 1 所示,主要包括 3 个步骤。第一是中文新闻评论情感词产生,通过新闻评论语料词频统计和基础情感词典进行比较分析得到。基础情感词典由大连理工大学信息检索研究室的情感词汇本体^[11]、HowNet 情感词典和台湾大学简体中文情感极性词典(NTUSD)构成。第二是中文新闻评论种子情感词产生,方法是在第一步产生的情感词的基础上,利用情感词汇本体中情感词的强度和人工判定相结合的方法,选取部分正面和反面情感强度最强的情感词作为种子词。第三是计算中文新闻评论情感的极性和强度,利用种子词、中文新闻评论语料以及本文提出的方法,进行情感词极性判定和强度计算。

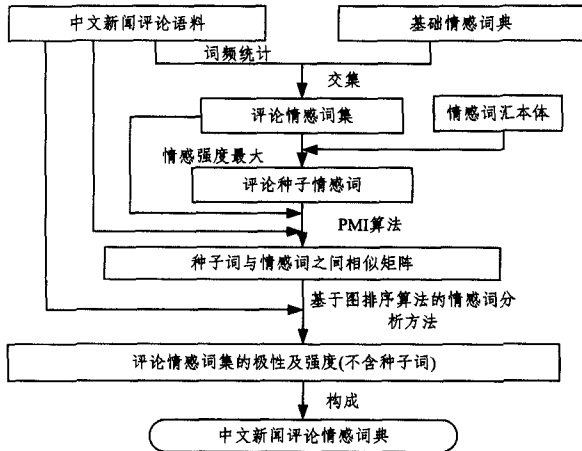


图 1 新闻评论情感词典构建基本思路框图

在情感词典构建过程中,首先设计程序将新闻评论语料下载后进行预处理,生成中文新闻评论语料集;然后进行词频统计等处理,为后续构建词典提供语料集。本文提出的确定评论情感词的极性及强度的方法的主要思路如下:首先选择种子情感词,并分配一个数值表示情感强度和极性,“正面”情感词为数值 1,“负面”情感词为数值 -1;对于待判定情感词的极性及强度,随机产生一个 -1 到 +1 的数值;将种子词与待判定词构成无向图,顶点表示词,边表示词之间的相似度,利用 PMI 算法在语料集中计算词的相似度;最后利用 PageRank 算法迭代计算待判定情感词的情感强度,算法收敛时得到最终的情感强度,并据此判别待判定词的情感强度和极性。

3 基于图排序算法的中文新闻评论情感词典构建方法

3.1 中文新闻评论情感词产生

研究中,利用我们设计的软件系统从凤凰网新闻资讯下载新闻评论,经过去除中性评论、垃圾评论、表情符号的转换

等预处理,形成 100 万条语料集。我们以这些语料集作为研究对象。评论语料集经分词和词频统计,去掉频率较低的词,产生新闻评论词集。将获得的新闻评论词集与基础情感词典中的词进行比较,词语的交集作为新闻评论情感词集,如“肮脏”、“安康”、“创新”等在评论词集和基础情感词典中均存在,则将它们归为评论情感词。经过上述处理共产生 555 个中文新闻评论情感词。

3.2 种子情感词的选择

种子词的选择对情感词典构成也很重要。在上述新闻评论情感词集的基础上,查询这些词在情感本体^[11]中的强度,将强度最大且在语料中出现频率较大的词作为候选种子情感词,例如“报仇”在情感本体中的情感强度为 9(最强),极性为负面,而在新闻语料中统计频率为 530 次,则将它作为候选种子情感词。经过上述处理,产生如表 1 所列的种子情感词,其中,“正面”40 个,“反面”40 个。

表 1 新闻评论种子情感词

正面种子情感词(40 个)
公道 英雄 公平 强大 幸福 公正 善良 自豪 英明 至高无上 创新 开心 创意 富裕 完美 称赞 弘扬 不得了 敏锐 智慧 团结 魄力 强健 钦佩 推进 谨慎 富强 喝彩 珍贵 廉政 英勇 安居乐业 歌颂 肃然起敬 富贵 孝敬 灿烂 光明正大 扬眉吐气 安稳 庆贺
负面种子情感词(40 个)
无法无天 悲剧 谣言 报仇 卑鄙 肮脏 猫腻 为所欲为 忍无可忍 残害 腐朽 肆无忌惮 草菅人命 见死不救 摧毁 忘恩负义 胡作非为 渎职 趁火打劫 急功近利 罪魁祸首 狼狽为奸 险恶 违规 卑鄙 令人发指 劣等 暴徒 残暴 贪赃枉法 雪上加霜 惨重 凌辱 暴戾 挑拨离间 污染 扭曲 肆无忌惮 令人发指 遗臭万年

3.3 基于语料的情感词语之间的相似度计算

本文提出的情感词典构建方法,需要计算 3.1 节产生的评论情感词(不含种子词)与种子情感词之间以及评论情感词之间的相似度,进而构建评论情感词与种子情感词的相似矩阵。我们的研究采用 PMI 算法来计算,如式(1)所示。经适当处理将 $PMI(word1, word2)$ 值作为情感词 $word1$ 与 $word2$ 之间的相似度。

$$PMI(word1, word2) = \log_2 \left(\frac{p(word1 \& word2)}{p(word1)p(word2)} \right) \quad (1)$$

其中, $p(word1)$ 和 $p(word2)$ 分别表示 $word1$ 和 $word2$ 在新闻评论语料库中出现的概率, $p(word1 \& word2)$ 表示 $word1$ 和 $word2$ 在语料中同时出现的概率。

3.4 基于 PageRank 模型的情感词极性判断和强度计算算法

(1) 算法描述和符号约定

本文提出的算法主要判定 3.2 节产生的新闻评论情感词极性并计算其强度,正面情感强度用 0~1 之间的值表示,负面情感词强度用 -1~0 之间的值表示。种子情感词集合向量为 $S = \{s_1, s_2, \dots, s_n\}$, 其情感极性向量为 $Y = \{y_1, y_2, \dots, y_n\}$, 当情感词属于正面情感词时, $y_i = 1$, 反之, $y_i = -1$ 。新闻评论情感词(不含种子情感词)向量为 $W = \{w_1, w_2, \dots, w_m\}$, 其情感极性及强度向量为 $P = \{p_1, p_2, \dots, p_m\}$ 。定义图 $G = \langle N, M \rangle$, $|N| = |S| + |W|$, 其中 N 为 G 中的结点集合(结点由全部情感词构成), $|S|$ 为种子情感词数, $|W|$ 为新闻评论情感词数, $|W| \times |N|$ 连接矩阵 M 描述结点间的无向图链接关系, M_{ij} 为结点 i 与结点 j 间的语义相似度, M 可分解为 $|W| \times |S|$ 的子矩阵 U 和 $|W| \times |W|$ 的子矩阵 V 两部分, U_{ij} 为新闻评论情感词 i 和种子情感词 j 之间的语义相似度, 由 3.3 节介绍的方法来计算相似度, 种子词将正面词放前面,

负面词放后面; V_{ij} 为新闻评论情感词 i 和新闻评论情感词 j 之间的语义相似度, 也由 3.3 节介绍的方法来计算相似度。根据 PageRank 算法的思想, 利用式(2)迭代计算新闻评论情感词情感强度及极性向量 P 。

$$P^{(t)} = (1-\beta)UY + \beta VP^{(t-1)} \quad (2)$$

式中, t 表示迭代次数, β 表示对 P 的贡献大小。为保证算法收敛, 算法每迭代一次都需要将 P 归一化。

(2) 算法

输入: $S, Y, W, P, K, \beta, \alpha, \gamma$

输出: w_i 和 p_i 待测词的情感极性及其强度(权值), P

算法:

Step 0 初始化 $P^{(0)} = 0, K = 40, \beta = 0.2, \alpha = 0.000001, \gamma = 0.01, t = 1, T = 500$ 。//表示迭代次数变量, T 表示最大迭代次数。

Step 1 构建矩阵 M 。其中, 连接权重按 3.3 节介绍的方法计算。正面种子词放前面, 负面种子词放后面, 新闻评论情感词(不含种子词)种子词对数量的选取见 4.1 节。根据矩阵 M , 产生矩阵 U 和矩阵 V 。

Step 2 矩阵 V 修改。按下面步骤修改矩阵 V 。

(1) 计算 $\max K(V_i^*)$, $1 \leq i \leq |W|$ 。 $\max K(V_i^*)$ 为 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。

(2) $\forall i, j, \exists V_{i,j} \leq \max K(V_i^*)$ 时, $V_{i,j} = 0$; 否则, $V_{i,j}$ 保持不变。

Step 3 归一化。

(1) 归一化 Y 。对 Y 进行归一化, 以保证 Y 中正面情感词分值和为 +1, 负面情感词分值和为 -1。公式如式(3)所示。

$$y_i = \begin{cases} -\frac{y_i}{\sum_{i=1}^{|S|} y_i}, & y_i < 0 \\ \frac{y_i}{\sum_{i=1}^{|S|} y_i}, & y_i > 0 \end{cases} \quad i=1, 2, \dots, |S| \quad (3)$$

(2) 归一化 U 和 V 。按照式(4)对 U 和 V 进行联合归一化。

$$U_{ij} = \frac{U_{ij}}{\sum_{j=1}^{|S|} U_{ij} + \sum_{j=1}^{|W|} V_{ij}} \quad (4)$$

$$V_{ij} = \frac{V_{ij}}{\sum_{j=1}^{|S|} U_{ij} + \sum_{j=1}^{|W|} V_{ij}}$$

Step 4 按式(1)计算 $P^{(t)}$ 。

Step 5 归一化 $P^{(t)}$ 。按式(5)进行归一化。

$$p_i^{(t)} = \begin{cases} -\frac{p_i^{(t)}}{\sum_{i=1}^{|W|} p_i^{(t)}}, & p_i^{(t)} < 0 \\ \frac{p_i^{(t)}}{\sum_{i=1}^{|W|} p_i^{(t)}}, & p_i^{(t)} > 0 \end{cases} \quad i=1, 2, \dots, |W| \quad (5)$$

Step 6 计算误差平方和。

$$\Delta = \sum_{i=1}^{|W|} (p_i^{(t)} - p_i^{(t-1)})^2 \quad (6)$$

Step 7 算法结束条件判定。

若 $\Delta < \alpha$ 或 $t \geq T$, 算法转 Step 8, 否则, $t = t + 1$, 算法转 Step 4。

Step 8 算法输出。

(1) 根据 $P^{(t)}$, 计算情感词 w_i 的极性并作出判断。

$$p_i = \begin{cases} \text{正面}, & p_i > \gamma \\ \text{负面}, & p_i < \gamma \end{cases} \quad i=1, 2, \dots, |W| \quad (7)$$

式中, p_i 的绝对值大小表示词 i 正反面强度。

(2) 输出其它参数 t, Δ

Step 9 算法结束。

4 实验分析

为了验证提出的方法的有效性, 拟从情感词判定准确性和基于构建的新闻评论情感词典分类性能两个方面进行实验。研究中, 设计软件从凤凰网新闻资讯网(<http://news.ifeng.com/>)采集了社会、资讯、文化等热门新闻的评论, 经过去除中性评论、垃圾评论、表情符号的转换、网络情感词的转换后, 语料集共有 100 万余条新闻评论。选择 4500 条评论进行“正面”和“负面”两个情感类别的人工标注。

4.1 新闻情感词集判定准确性实验与分析

在 3.2 节介绍中, 已产生 555 个新闻评论情感词, 扣除 80 个种子词后的评论情感词数量为 475 个。通过基础情感对比和人工确认, 这 475 个情感词中有“正面”184 个、“负面”291 个, 这些情感词作为待判定词进行实验。实验中采用的种子情感词“正面”和“负面”各 40 个(见表 1)。将种子情感词分成 3 种情况, 即 20 组对、30 组对和 40 组对, 来分析 K 和 β 的不同值对算法性能的影响。表 2 和表 3 是实验的结果。在表 2 和表 3 中, sw 表示种子词类别, $G2$ 表示 20 组对种子词, $G3$ 表示 30 组对种子词, $G4$ 表示 40 组对种子词; acc 表示算法对评论情感词判断的准确率。

表 2 β 参数影响实验结果($K=80$)

sw	β	acc(%)	sw	β	acc(%)	sw	β	acc(%)
G4	0.2	90.5	G3	0.2	81.89	G2	0.2	81.26
	0.4	92.63		0.4	90.10		0.4	82.10
	0.6	98.52		0.6	95.78		0.6	73.68
	0.8	98.95		0.8	96.84		0.8	67.36

从表 2 可以看出, 当种子词选取数逐渐减少时, 准确率(acc)下降趋势较明显; 也可以看到, 当 β 取值适当增大时, 准确率有较大上升。当然 β 不能无限增大, 通常的情况下取 $\beta < 0.5$ 为宜。

表 3 K 参数影响实验结果($\beta=0.4$)

sw	K	acc(%)	sw	K	acc(%)	sw	K	acc(%)
G4	10	75.78	G3	10	65.26	G2	10	58.94
	40	96.82		40	95.36		40	65.26
	80	98.73		80	83.57		80	72.84
	120	98.52		120	80.42		120	74.94
	250	73.68		250	62.94		250	60.02
	350	63.15		350	52.63		350	58.10

从表 3 可以看出, 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。 K 值取 80~120 较为适合。

4.2 新闻评论分类实验与分析

本实验采用作者所在研究小组前期提出的基于情感词典和朴素贝叶斯中文文本情感分类方法^[12], 对已标注情感类别的新闻评论语料进行分类实验。实验中, 在 $K=100, \beta=3$, 种子词 $G4$ 的条件下获得评论情感词典。已标注 4500 条新闻评

(下转第 80 页)

的概率估计为基础,引入改进的调度策略,建立作业与实时节点负载之间的协作关系,自适应调节属性概率,调度最适合执行的优先任务。最后通过实验对改进的调度算法进行测试,并与经典调度算法进行比较。实验结果显示,将地震资料数据进行静态化划分后,对划分后的作业结合模糊聚类优化策略,通过改进的贝叶斯分类调度算法实施动态调度,达到了较高的计算效率。

参 考 文 献

[1] 赵春燕. 云环境下作业调度算法研究与实现[D]. 北京:北京交通大学,2009

[2] 罗银河,刘江平,俞国柱. 叠前深度偏移述评[J]. 物探与化探,2007,28(6):540-546

[3] Dean J, Ghemawat S. Mapreduce: Simplified Data Processing on Large Clusters[J]. Communications of the ACM, 2008, 1: 107-109

[4] 李文娟,张启飞. 基于模糊聚类的云任务调度算法[J]. 通信学报,2012,33(3):146-153

[5] 李柏年. 一种改进的模糊 C-均值算法[J]. 计算机应用与软件,

2009,25(6):21-25

[6] 夏伟. Hadoop 平台下的作业调度算法研究与改进[D]. 广州:华南理工大学,2010

[7] 余正祥. 基于学习方式对 Hadoop 作业调度的改进研究[J]. 计算机科学,2012,39(6):220-224

[8] Hand D J. 机器学习十大算法:朴素贝叶斯[M]. Taylor & Francis Group, LLC, 2009:163-178

[9] Zhang Shu-fen, Zhang Shuai, Chen Xue-bin, et al. Analysis and Research of Cloud Computing System Instance[J]. 2010 Second International Conference on Future Networks, 2010, 60:88-92

[10] Randles M, Lamb D, Taleb-Bendiab A. A comparative study into distributed load balancing algorithms for cloud computing[C]// Proc of the 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, Fukuoka, Japan, 2011:551-556

[11] Hu Jin-hua, Gu Jian-hua, Sun Guo-fei. A scheduling strategy on load balancing of virtual machine resources in cloud computing environment[C]// Proc of the 3rd International Symposium on Parallel Architectures, Algorithms and Programming, Liaoning, China, 2010:89-96

(上接第 69 页)

论,“正面”2100 条,“负面”2400 条,采用 k-fold($k=10$)交叉验证的方法来验证,同时使用不同的情感词典进行对比实验,实验结果如表 4 所列。

表 4 新闻评论情感分类实验结果

使用的情感词典	正面评论 准确率	负面评论 准确率	平均准确率
本文构建的情感词典	80.95%	87.50%	84.44%
情感词汇本体	76.19%	81.25%	78.89%
NTUSD	69.04%	72.91%	71.10%
HowNet 情感词典	56.51%	68.13%	66.73%

从表 4 可以看出,本文所构建的情感词典分类准确率较其它情感词典高,表明本文提出的方法具有有效性;同时,本文构建的情感词典在训练分类器时所需时间最短。

结束语 本文针对网络新闻评论,提出了新闻评论专题领域情感词典构建方法。这种方法利用语料库和基础情感词典产生专题领域情感词集,在情感词集的基础上,获取种子情感词典,然后利用图排序模型原理,通过迭代运算,获得专题领域情感词的极性和强度,进而构建情感词典。文章通过实验和结果分析,验证了提出的方法和构建的词典的有效性。本文提出的方法也可以用于其它专题领域情感词典的构建。如何自动确定本文提出的方法的各参数,我们将做进一步探讨。

参 考 文 献

[1] 杜伟夫. 文本倾向性分析中的情感词典构建技术研究[D]. 哈尔滨:哈尔滨工业大学,2010

[2] 许静芳,李星,李粤. 信息检索中主题式词典的构建方法[J]. 计算机工程,2005(21):153-155

[3] 柳位平,朱艳辉,栗春亮,等. 中文基础情感词典构建方法研究[J]. 计算机应用,2009(10):2875-2877

[4] Xu G E, Meng Xin-fan, Wang Hou-feng. Build Chinese emotion-

lexicons using a graph-based algorithm and multiple resources [C]// Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, PA: Association for Computational Linguistics, 2010:1209-1217

[5] Kim S M, Hovy E. Identifying and analyzing judgment opinions [C]// Proceedings of the Main Conference on Human Language-Technology Conference of the North American Chapter of the Association of Computational Linguistics, Stroudsburg, PA: Association for Computational Linguistics, 2006:200-207

[6] Kim S M, Hovy E. Automatic detection of opinion bearing word-sand sentences [C] // Proceedings of the Second International Joint Conference on Natural Language Processing, Jeju Island: [s. n], 2005:61-66

[7] Hatzivassiloglou V, Mckeown K. Predicting the semantic orientation of adjectives [C]// ACL-97: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain: [s. n], 1997:174-181

[8] Velikovich L, Blair-Goldensohn S, Hannan K, et al. The viability of Web-derived polarity lexicons [C]// Proceedings of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA: Association for Computational Linguistics, 2010:777-785

[9] Turney P, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4):315-346

[10] Brin S, Page L, Motwami R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford: Stanford University, 1999

[11] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报,2008,27(2):180-185

[12] 杨鼎,阳爱民. 一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J]. 计算机应用研究,2010,27(10):3737-3739