

基于权值算法的中文情感分析系统研究与实现

张昊旻, 石博莹, 刘栩宏

(北京邮电大学 软件学院, 北京 102209)

摘要: 为了能够更加准确地对语句结构进行划分、对语句表达的内容进行判断,提出了一种全新的基于权值的计算算法,在完善中文分词的基础上对语句进行情感分析。首先利用中文分词算法对句式结构进行分割,然后依据词性对词库进行扩展,词库对句中干扰词汇进行过滤,最后利用全新的权值计算算法对语句情感进行准确分析。经有效测试结果表明,情感分析准确率较高,并广泛适用于网络舆情分析等应用中。

关键词: 中文情感分析; 权值算法; 句式结构; 情感强度; 舆情

中图分类号: A31 **文献标志码:** A **文章编号:** 1001-3695(2012)12-4571-03

doi:10.3969/j.issn.1001-3695.2012.12.044

Study and implementation of Chinese language emotion analysis system based on weight algorithm

ZHANG Hao-min, SHI Bo-ying, LIU Xu-hong

(School of Software Engineering, Beijing University of Posts & Telecommunications, Beijing 102209, China)

Abstract: In order to separate the sentence structure and to judge the content of the statement more accurate, this paper put forward a new methodology of weight calculation based on the foundation of Chinese words segmentation and then analyzed the emotion of sentences. First, it adapted Chinese words segmentation algorithm to separate the sentence structure and then appended word dictionary according to the part of speech of words. Next, the methodology filtered disturbing words in sentences using word dictionary. Finally, the methodology took advantages of the new weight calculating algorithm to analyze the emotion of sentences accurately. Several validation tests approve that the result of the analysis has high accuracy and can be widely used in many applications such as public opinion analysis and so on.

Key words: Chinese emotion analysis; weight algorithm; structure of sentence; strength of emotion; public opinion

0 引言

如何从海量文本信息中提取出关键、有价值的文本信息为人们所用成为当前研究的重点。由于中文文本的特殊性,不能像英语等语言可以利用空格作为切词断点进行高效准确的分词,因此对中文分词的准确与否也将影响对内容进行情感分析的结果正确性。正是基于这样的研究背景,中文情感分析的研究价值就体现了出来,能够对中文语句段落进行情感分析将会有很大的应用价值。例如,中文情感分析就能帮助网络舆情监控系统的开发。或支持或反对、或理性或感性、或热情参与或冷眼旁观。总而言之,人们大相径庭的态度对事件的后续发展与走向起着至关重要的作用,而对这些态度的收集与分析可以极大地帮助企业或者个人把握社会的动态,以利用这些数据帮助领导者进行决策^[1]。因此可以看出中文情感分析的应用价值是十分可观的。

然而目前此方向研究领域,很多的研究者将目光放在某些专业领域内容中文情感分析的研究和实现。例如,目前市面上存在成型的系统是专门用于分析新闻领域。对特定领域进行情感分析的方法大多采用贝叶斯分类算法进行。贝叶斯是一种基于概率的学习算法,能够用来计算显式的假设概率,它基于假设的先验概率,给定假设下观察到不同数据的概率以及

观察到的数据本身^[2,3]。因此,需要有极大且准确的包含领域主体词的训练集作为支撑,产出相应领域内容的贝叶斯分类器,以进行领域内的中文文本情感分析。所以,在面对纷繁复杂的网络评论内容时,此算法就不能够适用了。因此,基于对此问题的分析与研究,本文提出一种基于权值的中文情感分析算法,目的就是为了能够应用于多领域的情感分析,增强其普遍应用价值。

本文所介绍的是利用词典以及句式词库自动化分析中文句式的特殊结构以及情感倾向词,并配合权值算法来代替传统人工判别或仅仅利用简单的统计方法进行网络舆情监控中文情感分析,有效提高了在普遍意义的网络环境中中文情感分析的效率以及准确度。

为了更加适应网络评论内容的广泛性和通用性,本算法放弃使用需要大量训练数据的贝叶斯分类算法来进行情感分析计算,而是精炼出中文语言的特点,采用了针对中文分词结果、中文特殊句式以及句子主张词的判断的中文情感权值算法来进行情感倾向值计算,使得利用该算法的文本分析系统在对通用领域的判别上更加准确^[4]。

1 中文分词分析与研究

智能化分析中文语言有一定的难度,因为汉语虽然有自己

收稿日期: 2012-04-26; 修回日期: 2012-05-30

作者简介: 张昊旻(1990-),男,北京人,本科生,主要研究方向为软件工程、自然语言处理(bluezebra.zhang@gmail.com);石博莹(1991-),女,北京人,本科生,主要研究方向为软件工程、自然语言处理;刘栩宏(1990-),女,四川岳池人,本科生,主要研究方向为人工智能、自然语言处理。

的一套语法结构,但是在实际运用中,如何进行分词,句子的句式影响,又会因为场景的不同而改变,所以想要相对准确地进行中文情感分析,就需要有准确的分词方法作支撑。而本系统采用的中文分词算法,是对中国科学院分词算法的改进。首先精简了很多在日常生活中很少使用到的词汇,并将主语词提取出来,以判断句子所表达的情感是不是由语句作者所提出的;除此之外,系统还将主张词汇加入词库之中,以明确主张词与情感词的分离;最后,系统将一些由单个词汇组成的组合词添加到了词库之内。这样做的好处是,既能明确公众主张,又能准确地分割句子结构,使得语句的情感得到最大程度的正确匹配^[5]。为了能尽早测试算法准确度,以上工作目前仅由人工手动调整。

首先分割主语词。网络评论的特点在于人们的言论以电子化的文档形式存在,而非日常生活中口口相传,这其中最重要的就是要确定语句所表达的情感是否真实来源于语句作者本人^[6]。如果在句子中出现了一些“他说”“XX 曾经说”这样的词汇,就很难确定语句的情感到底是作者本人,还是引用自原文作者。在本系统当中,有一个主语词词库,要被分析的句子会首先被放入主语词词库进行筛选,将与作者无关的词过滤掉以增加准确率^[7]。

除此之外,系统还涉及到过滤主张词。主张词的作用更多的是引出后文,对某一具体事件的态度或情感,而其本身是没有任何感情色彩的,因此这些主张词对情感分析是没有任何帮助的。系统根据大量的资料,人工将频繁出现在人们日常生活中的主张词进行了收集和整理,建立了相应的主张词词库。

第三个基础词汇部分是副词。普通的副词已经在以前的分词算法中有所体现,在此不作赘述。这里想要强调的是,笔者所在团队开发的分析系统着重考虑了一些组合副词,这些词通常是由一些基础的副词或者形容词组成,这些基础词汇通常具有自己的情感倾向,例如对“不得不”进行分词(表1)。

表1 中国科学院分词结果

Input	不得不	
Output	不得	不

通常情况下,以分词结果的词义进行情感分析时,结果会将两词分别当成表发负面情感的词汇。但是实际上,“不得不”表达的并不完全是负面的情感,而要取决于它之后的情感词汇。比如“我不得不承认,这场演出无与伦比。”这里的主情感词汇是“无与伦比”,而“不得不”并没有否定这种情感。再比如“我不得不承认,这场演出彻彻底底失败了。”这句话的主情感词是失败,很明显的负面词汇,而“不得不”也是加强了这种否定的语气。因此,通过添加组合副词扩大词库,可以有效地避免因为这类词汇所造成的误差。

2 句式结构的分析与研究

在进行完中文分词操作之后,句子已经被分割成很多不同的部分,而对情感分析造成干扰的无关项也已经被有效地过滤掉了。按照以往的经验,下面就可以对语句进行情感分析了。但是同时以往的教训也预示着,如果就这样匆匆忙忙地进行情感分析,结果会出现比较大的偏差,有时偏差甚至可能达到30%。经过分析与研究,笔者团队达成了共识,误差主要是因为旧有系统缺乏对中文句式的分析,而盲目地对分词结果进行情感分析。实际上,不仅仅是中文语句,所有的语句都存在句

子结构,如何能够彻底地分解句子结构,找出真正为语句情感作出贡献的成分,是亟待解决的问题,也是提高情感分析准确率的重中之重。经过大量的调查与总结,以下句式对于情感分析有着比较明显的意义^[8]。

2.1 联合复句

并列复句、递进复句这两种句式在情感分析过程中起到的作用是一致的,可以被归类为加感情词汇的情感程度。对于并列复句,通常是以A+B的形式出现。A、B两个复句之间所描述的事物本身就是处于同样的地位,两者之间没有任何顺承或是相互否认的关系,因此可以将A、B两个复句的情感程度进行叠加。而对于递进复句,句式一般为A+A⁺,即后半复句所表达的事物通常是对第一复句事物更详细地描述,而在表达情感时,第二复句所蕴涵的情感要比第一复句更加深刻和强烈,因此在进行情感分析这两种句式的句子时,可以两个复句中的情感进行叠加,以得到整句的情感色彩^[9]。

添加入词库的,常见的表达并列关系的连接词有“和”“跟”“与”“一边……一边……”“又……又……”“先……再……”“既……又……”“并且……”等。

添加入词库的,常见的表达递进关系的连接词有“更……”“甚至……”“不但……而且……”“何况……”“不但……”等。

2.2 偏正复句

2.2.1 转折复句

转折句的句式结构可以表达为A…转折词B…。转折复句是句子结构中比较特殊的一种句式,无论前文如何强调事物的特点或是表达强烈的情感,都会被转折词所否定,因此在分析时需要尤其注意的是转折词的捕捉^[10]。当系统在进行分词之后会重新遍历等待分析的语句,一旦捕捉到转折词,系统就会自动将该词语之前已经分析过的句子的情感取反,直接得到该句所要表达的感情色彩。同理,如果在某一个段落中间,系统捕捉到了转折词汇,转折词之前的段落情感将被直接取反。经过大量的测试,这种判断方法的误差保持在3%~5%之内,极大地简化了情感分析的过程,有效地提高了算法的效率。

添加入词库的、常见的表达转折关系的连词有“却”“但是”“然而”“而”“反而”“偏偏”“只是”“不过”“至于”“不料”“岂知”等。

2.2.2 目的复句、因果复句

目的复句和因果复句的情感一般只出现在多个复句中的其中一部分,而其他部分则是为叙述这种事物、表达这种情感所做的铺垫。因此与转折句式不同的是,在分析目的复句和因果复句时,在捕捉到相关连词之后,系统需要抛弃之前语句的态度,而重点分析该连词之后出现的情感词的色彩^[11]。虽然在表达目的与结果之前,句子可能会叠加很多情感词汇,但是这些词汇基本上都是为了表达作者的最终看法所存在,因此可以看做是一种铺垫,而没有实际的意义。直接准确地捕捉到目的词汇后面的情感词汇,大大提高了分析的准确程度^[12]。

添加入词库的,常见的表达目的关系的连词有“以”“以便”“为了”“使得”等。

添加入词库的,常见的表达因果关系的连词有“原来”“由于”“因此”“(之)所以”“以致”“是故”“因为”“于是”“就”等。

2.2.3 让步复句

让步复句的感情色彩也是取决于让步连词之后的复句,对于句子本身来说,单独将让步句式分为一类意义并不大,但是对于一个段落,让步复句的存在对于情感色彩的强烈程度却有着很大的影响。因为系统使用的情感分析算法采用的是权值叠加的算法,对于其他句式而言,确定了某些感情词的色泽和程度之后,系统就可以进行叠加了。然而对于让步句式,在情感程度出现之后,因为句式的缘故,程度将会有所削弱,而这是其他句式所不具有的特点。因此,在利用分词进行句式判断之后,在情感分析让步句式时,系统会自动在情感程度上有所降低。

添加入词库的,常见的表达让步关系的连词有“就算”“哪怕”“即使”“纵然”“无论”“尽管”等。

3 情感分析的研究与改进

改进后的中文分词算法已经为情感分析提供了极大的便利。系统需要做的就是给强烈程度不同的情感词赋予以不同的权值,然后再利用算法进行权值的加和^[13]。

3.1 为不同情感词赋值

汉语中不同词语在表达情感上会有不同的情感强烈程度,为此需要人工为词库中的感情词按情感强烈程度添加权值,以便区分情感等级。

例如“高兴”“欢呼雀跃”“手舞足蹈”“欣喜若狂”,它们的权值分别为+1、+3、+3、+5。“高兴”表达的是正面情感,因此用“+”来表示,但是又因为“高兴”只是很轻微地表达了情感,因此权值仅为1。同理,对于表达负面情感的词汇,首先权值为负值,其次按照情感的强烈程度赋值。

3.2 权值计算

本文所采用的权值计算方法是对加权平均算法的改善,目的是为了能够更好地与情感分析相匹配。改善过后的加权平均算法公式为

$$\bar{E} = \frac{\sum_{i=1}^{N_p} wp_i + \sum_{j=1}^{N_n} wp_j}{N_p + N_n} \quad (1)$$

其中: N_p 、 N_n 分别代表表达正面情感和负面情感的词汇数目; wp_i 、 wp_j 分别代表正面情感词汇和负面情感词汇的权值(含符号)。

首先,在进行情感分析时要考虑到不同类型词汇数量对于句子的影响。对于一般句式来说,包含负面情感词多的句子一般会表达比较负面的情绪,而当正面情感词占据句子的主体地位时,句子的情感一般都是偏于正面的,因此情感词的数量有较大影响。其次,感情词的强烈程度及权值会与情感词数量共同作用,因为即使某类情感词出现的次数很多,它们的作用也会因为一些相反情感词的权值较大而失去其功效。综合了这两种考虑,用式(1)来计算整句话的情感是比较客观且全面的。

4 其他改进

除了以上对于算法的改进,本文还对中国科学院分词词库进行了扩充。由于本文阐述的系统重点分析的是互联网上主流论坛、贴吧中网民留言的感情与态度,在扩充词库时,主要考虑的是与当代社会相适应,因此本文利用了搜狗输入法所提供的网络热词库,如“顶”“赞”等类似词汇,使得系统更加与时俱进。

5 实验结果

为了能展现整个分析过程,采用例句1“我不是不想顶你,只是对我也没好处。”和例句2“我无所谓,但这样做道德败坏,不值得提倡的。”作为分析说明。

第一步进行中文分词处理。例句1和2的处理结果如表2所示。

表2 例句1和2中文分词结果

例句1	Input	我不是不想顶你,只是对我也没好处。
	Output	我 不是 不 想 顶 你,只是 对 我 也 没好处。
例句2	Input	我无所谓,但这样做道德败坏,不值得提倡的。
	Output	我 无所谓,但 这 样 做 道 德 败 坏,不 值 得 提 倡 的。

第二步则是对其进行中文情感分析计算。为了能够清楚说明分析过程,只对例句1进行分析,例句2不再赘述,分析结果如表3所示。

表3 例句1中文情感分析步骤

步骤	处理结果
1) 句式分析	只是一→转折词汇
2) 有用词汇	对 我 也 没好处
3) 停用词	对 也
4) 主张词	我
5) 可计算词汇	没好处
6) 分析结果	oppose
7) 转折词前分析	依据步骤2)~5)分析可得结果 support
8) 取反	将转折词前分析取反得 oppose
9) 比对	步骤6)结果与步骤8)结果相同
10) 结果	本句为 oppose

通过对所输入的例句1进行以上逻辑步骤的分析和处理,将能够较为准确地判断其内容的情感倾向性,遂将例句2也放入本系统进行中文情感分析,其结果同样为 oppose。Oppose 的意思是:此句话的中文情感分析结果为反对态度。针对例句1和2的系统测试情感分析结果如图1所示。



图1 例句1和2情感分析结果

6 结束语

本文采用的是基于中文分词、分析语句结构与权值计算的方法来判断语句的情感。在实际的操作方面,使用了中国科学院的分词算法,对大量无用实体词进行了筛选,并且加入了最新流行词汇,尽可能地使得网络舆情分析准确和快速。

未来需要改进的地方有:

- 改进分词算法,进一步提高分词速度,提高系统效率。
- 改进语句中连词与词库的匹配算法,提高匹配效率。
- 继续扩充词库,使得系统可以与搜狗词库同步,扩大词汇量。
- 在权值的分配上更加合理,精化权值算法,使情感程度的权值计算更加精确。

- [3] BERLEANT D, KUIPERS B J. Qualitative and quantitative simulation; bridging the gap[J]. *Artificial Intelligence*, 1997, 95(2): 215-255.
- [4] DANIEL B. The use of partial quantitative information with qualitative reasoning[D]. Austin: University of Texas at Austin, 1991.
- [5] 刘丙杰, 胡昌华, 李海亮. 一种改进的半定量仿真算法[J]. *系统仿真学报*, 2007, 19(5): 963-965.
- [6] SHEN Qiang, LEITCH R. Extending qualitative simulation by the use of fuzzy sets[J]. *IEEE Trans on Systems, Man, and Cybernetics*, 1992, 22(2): 1001-1008.
- [7] SHEN Qiang, LEITCH Roy. Fuzzy qualitative simulation[J]. *IEEE Trans on Systems, Man, and Cybernetics*, 1993, 23(4): 1038-1061.
- [8] LEITCH R, SHEN Qiang. Qualitativeness does not imply fuzziness[J]. *IEEE Trans on Systems, Man, and Cybernetics*, 1994, 24(2): 1257-1261.
- [9] BELLAZZI R, GUGLIELMANN R, IRONI L. Learning from biomedical time series through the integration of qualitative models and fuzzy systems[J]. *Artificial Intelligence in Medicine*, 2001, 21(3): 215-220.
- [10] 王东锋, 陈英武. 模糊定性建模与仿真集成实现[J]. *系统工程理论与实践*, 2002, 22(3): 119-124.
- [11] HILDE D. Qualitative simulation and related approaches for the analysis of dynamic systems[J]. *The Knowledge Engineering Review*, 2004, 19(2): 93-132.
- [12] 刘丙杰, 胡昌华. 基于高斯隶属函数的模糊定性仿真[J]. *系统工程与电子技术*, 2006, 28(7): 1098-1102.
- [13] BROWN D J, LIU Hong-hai. An extension to fuzzy qualitative trigonometry and its application to robot kinematics[C]//Proc of IEEE International Conference on Fuzzy Systems. Piscataway: IEEE Press, 2006: 1111-1118.
- [14] 夏文俊. 面向概念设计的定性建模技术研究[D]. 武汉: 华中科技大学, 2011.
- [15] 李文伟. 定性仿真方法发展与应用研究[J]. *系统仿真技术*, 2008, 4(2): 71-74.
- [16] HUANG Yuan-liang, CHEN Zong-hai, GUI Wang-shen. Grey qualitative simulation[J]. *The Journal of Grey System*, 2004, 16(1): 5-20.
- [17] 黄元亮. 灰色定性仿真基础研究[D]. 合肥: 中国科学技术大学, 2004.
- [18] 黄元亮, 潘冰. 基于灰色模型的系统半定量关联矩阵[J]. *系统仿真学报*, 2007, 19(3): 474-477.
- [19] HUANG Yuan-liang. Qualitative-quantitative system identification based on grey model[J]. *The Journal of Grey System*, 2005(2): 111-118.
- [20] KUIPERS B J. Qualitative simulation[J]. *Artificial Intelligence*, 1986(29): 289-338.
- [21] MARCO P, BERNHARD R, REINHOLD W. A specialized computer architecture for QSIM[J]. *IEEE Intelligent System*, 2000, 15(2): 62-68.
- [22] 梁昌勇, 杨善林, 黄梯云. QSIM 算法的序列因果关系约束和目标搜寻策略研究[J]. *计算机应用研究*, 2001, 18(1): 21-24.
- [23] 胡斌, 殷芳芳. 集成 CA 与 QSIM 的非正式组织群体行为演化的定性模拟[J]. *中国管理科学*, 2005(5): 130-136.
- [24] ZHANG He, BARBARA K, ROSS J. Planning software project success with semi-quantitative reasoning[C]//Proc of Australian Software Engineering Conference. Piscataway: IEEE Computer Society, 2007: 369-378.
- [25] SHAO Chen-xi. Quantitative and qualitative forecasting applied to supply-chain inventory[C]//Proc of the 2nd International Conference on Multimedia and Information Technology. Piscataway: IEEE Computer Society, 2010.
- [26] YANG Hua, HU Bin, ZHANG Liu-pei. Analysis for the quality risk evolution of highway engineering construction based on qualitative simulation[C]//Proc of International Conference on Computer Science and Service System. Piscataway: IEEE Computer Society, 2011: 212-215.
- [27] 陈培友, 李一军, 高太光. 基于 Kuipers 定性仿真算法的多议题谈判模型研究[J]. *中国管理科学*, 2011, 19(3): 158-165.
- [28] 付斌, 李道国, 王慕快. 云模型研究的回顾与展望[J]. *计算机应用研究*, 2011, 28(2): 420-426.
- [29] 邱凯昌. 空间数据发掘与知识发现[M]. 武汉: 武汉大学出版社, 2001: 43.

(上接第 4573 页)

参考文献:

- [1] 王翠波. 基于文本情感挖掘的企业技术竞争情报采集模型研究[J]. *图书情报工作*, 2010(14): 75-78.
- [2] Phinecos. 基于朴素贝叶斯分类器的文本分类算法(上). [EB/OL]. (2008-10-21) [2012-04-26]. <http://www.cnblogs.com/phinecos/archive/2008/10/21/1315948.html>.
- [3] Phinecos. 基于朴素贝叶斯分类器的文本分类算法(下). [EB/OL]. (2008-10-21) [2012-04-26]. <http://www.cnblogs.com/phinecos/archive/2008/10/21/1316044.html>.
- [4] KIM S M, HOVY E. Determining the sentiment of opinions[C]//Proc of the 20th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004: 1367-1373.
- [5] NASUKAWA T, YI J. Sentiment analysis: capturing favorability using natural language processing[C]//Proc of the 2nd International Conference on Knowledge Capture. New York: ACM Press, 2003: 70-77.
- [6] WILSON T, HOFFMANN P, SOMASUNDARAN S, et al. Opinion-Finder: a system for subjectivity analysis[C]//Proc of HLT/EMNLP on Interactive Demonstrations. Stroudsburg: Association for Computational Linguistics, 2005: 34-35.
- [7] HU Min-qing, LIU Bing. Mining and summarizing customer reviews[C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 168-177.
- [8] 曹逢甫. 汉语的句子与子句结构[M]. 北京: 北京语言大学出版社, 2005.
- [9] 陈中干. 现代汉语复句研究[M]. 北京: 语文出版社, 1998.
- [10] 郭志良. 现代汉语转折词的研究[M]. 北京: 北京语言文化大学出版社, 1999.
- [11] 申小龙. 汉语句型研究[M]. 海口: 海南人民出版社, 1989.
- [12] 刘建波, 杨峰. 面向舆情分析的短文本频繁模式聚类算法[J]. *北京电子科技学院学报*, 2010, 18(4): 6-11.
- [13] 龚静, 田小梅. 基于文本表示的特征项权值计算方法[J]. *电脑开发与应用*, 2008, 21(2): 46-48.