基于中文文本分析的微博情感地图的制作①

郭义超, 樊 红

(武汉大学, 武汉 430079)

摘 要: 自 web 进入 2.0 时代以来, 互联网社交信息爆炸式地融入了人民生活, 对海量社交网络信息的分析成为文本分析领域的一个重要研究方向. 本文通过整理情感词典, 制定语义规则, 分析评测中文微博的情感色彩并与 GIS 空间分析方法相结合绘制出了情感地图. 试图以客观的评价手段, 对主观情绪进行科学计量化描述, 并以地图为载体进行直观表达. 论文将微博情感分析结果作为公民幸福指数的评价参考, 同时, 将地理信息科学与传统的情感分析相结合制作出情感地图, 能够为国民幸福指数宏观评价及其空间分布特征提供更直观的展示和参考.

关键词: 微博; 情感分析; 地图

Approach to Make Mood Maps Based on Sentiment Analysis of Chinese Micro-Blog

GUO Yi-Chao, FAN Hong

(Wuhan University, Wuhan 430079, China)

Abstract: Since the time of web2.0, social network information of internet has increasingly integrated into people's life. The analysis of the vast social network information has become an important research direction in the field of text analysis. In this article, through establishing emotional dictionary and semantic rules, the emotional colors of Chinese micro-blog are analyzed and calculated. Then the emotion map combined with GIS spatial analysis method gives a scientific and quantitative description of subjective emotion by means of objective evaluation and the emotion revealed on the map. This article which takes the result of micro-blog sentiment analysis as evaluation index to evaluate the Gross National Happiness, analyzes the meaning of evaluating the happiness with mood map created on the approach which combines the GIS with the traditional sentiment analysis.

Key words: micro-blog; sentiment analysis; map

随着互联网技术以及移动技术在人民生活中的广泛普及, 人们的现实社交也逐渐虚拟化、平台化、移动化, 伴随的社交网络信息也呈现出爆炸式增长的态势. 微博作为一种重要的社交网络表现形式, 凭借联通全民的特点, 自诞生之初便迅速积累了数以亿计的用户. 以新浪微博为例, 每天都有数千万活跃用户在这个巨大的网络环境向世界表达着自身的情感. 并且移动定位的普及给微博附加了位置信息, 让本就迷人的社交网络增添了一份动感. 微博正在从各个方面渗透并影响人们的生活, 包括大量信息传播更快的信息发现与世界的连接等[1].

如此巨大的用户基础以及文本数据, 正吸引着越来越多的学者和机构对微博进行着各式各样的学术研究, 情感分析便是社交网络研究中开展最为广泛的一项. 情感分析是利用主、客观评价手段对文本、语音等语言表述形式中隐含的情感、态度进行挖掘, 得出文本、语音中表达的积极、消极或中立这样的极性指标的研究过程.

由此可见, 文本情感分析过程是指文本情感极性判别和文本情感强度的评价[2]. 因此跟分类方法相似, 文本情感分析也主要监督分类和非监督分类两种. 在情感分析领域, 监督分类是以事先经过专家标注好极

① 基金项目:国家自然科学基金(41471323)

收稿时间:2016-05-17;收到修改稿时间:2016-06-20 [doi:10.15888/j.cnki.csa.005594]

性的文本作为训练样本, 利用机器学习中的监督分类方法, 判断待分类的文本情感极性的分类方法. 这种分类方法的好处是能够自动分析语句中特征词汇中的语义联系, 利用合适的机器学习算法, 如人工神经网络, 对长文本能够达到很高的计算精度. 但是监督分类中分类器的训练需要一定数量经过标注的训练样本[3]. 并且需要对训练样本保证较高的标注精度和语义类型覆盖度, 并且微博限于篇幅和主题的原因, 情感特征往往偏少并且不规则, 因而监督分类在微博情感分析中往往难以发挥其高精度的价值. 非监督分类则无需提供事先标注好的训练样本, 这种不依赖注解数据的方法可以用来解决监督分类对于训练样本依赖性的问题[4].

自监督分类被引入文本分类起, 大量的监督学习方法也开始逐步被融合进来. 从朴素贝叶斯到支持向量机再到深度学习方法如卷积神经网络, 人们试图使在其它分类领域, 如图像识别中已获成功的方法引入文本分类, 以获取更好的分类结果. 事实也取得了比较积极的进步, 但文本分类有属于自己的特征, 如文本特征向量的前后关联性, 反义词向量引起情感极性转移等问题与传统的分类之间的差别, 使得这些新兴的分类方法难以在文本分类领域获得极大的成功. 当然, 在此基础上, 人们也应用了一些更加适合于文本或是语音特征的深度学习方法, 如递归神经网络, 从而获得了更高的分类精度. 另外还有学者将少量训练样本和大量未标注样本相结合, 并辅以主动学习, 迁移学习等机器学习方法形成半监督分类, 用于情感分类, 也取得了较好的成果.

文本情感分析中的非监督分类往往是基于情感词汇和词频统计的方法. 通过对待检测的文本进行词汇分割, 并标注出词汇的情感极性, 然后基于线性加权的算法计算文本的整体极性. 而情感词汇的整理在文本研究中已经取得了广泛进展, 获取途径多样且简单, 并且结合定义良好的规则匹配算法能够达到较好的分类精度, 因此适合于微博这种短文本的情感分类.

本文结合了非监督学习的优点, 利用覆盖全面的情感词典, 使用自定义的语义规则匹配算法, 对附有坐标信息的微博文本进行了情感分析. 对分析结果进行空间插值, 以此作为国民幸福指数的评价指标, 在地图上进行展现, 成功制作了精度较高且形象直观的情感地图, 对国民幸福指数及其空间分布进行了直观的可视化展示和表达.

1 相关工作

情感地图的制作通常分为情感分析和地图制作两个阶段. 情感分析涉及的技术包括文本分割, 词性判断, 情感特征识别和结果处理的过程. 情感分析的预处理工作主要包括语义规则的制定和进行规则匹配之前对微博文本的处理. 文本处理技术包括分词、词性标注、句法分析等自然语言处理技术[5].

1.1 文本分割

文本分割是将语义连续的语句、段落划分为词汇集合的过程. 词汇是许多自然语言处理系统的重要组成单元[6]. 因为词性判断的处理对象是词汇而非句子, 所以文本分割可以视为词性标注的预处理过程. 本文采用了 word 分词组件来实现这一过程, word 分词器是基于 JAVA 语言的一款分布式分词工具, 具有用户配置多样化, 模式灵活等特征. 由于微博中的新词汇层出不穷, 同时存在多种不规则词汇, 因此在微博情感分析中通常需要根据现有的微博语言环境自定义大量新词汇以及词组. 在这种背景下, 能够提供自定义词汇拓展, 自定义词汇优先级和词性标注的 word 分词组件自然更适合作为中文文本分析中的文本分割工具.

1.2 词性判断

词性判断主要是基于情感词典的词汇匹配, 带有词性标签的词汇所构成的特征对于文本分类至为重要[7]. 因此在词性判断阶段的准备工作主要是情感词典的准备和扩充. 本文整合了目前公开的多个情感极性词汇库, 包括台湾大学整理的 NTUSD 情感词典、知网情感词典、手工整理的微博表情符号词典, 同时加入了年度网络用语. 这样做的目的是尽可能多的涵盖微博文本中的情感特征, 消除网络词汇与传统词汇的异义性.

这样经过文本分割和词性判断后的微博文本就由一组有序汉语词汇变为了一个计算机可处理的由简单数字组成的一维数组.

2 基于语义规则匹配的情感分析

通过建立合理且尽可能覆盖语言模型的语义规则, 并将其应用到特征词数组中的过程, 即为基于语义规则的情感分析的情感极性计算过程.

2.1 基本定义和符号

情感分析的基本单元, 也是语义模型中的属性(property)是情感词汇[8], 如高兴, 伤心, 难过等, 以符号 S 表示. 情感词汇按积极和消极类型分别以 PS 和

NS 表示; 另外语义模型中的其他属性还包括程度副词, 如很, 非常, 极其等, 以符号 adv 表示. 否定词, 如不是, 并非等, 以 neg 表示.

2.2 语义规则

语义模型的计算规则和描述如表 1 所示, 如序列 1 表示积极情感词汇连续分布则进行加法运算, 具体分值为两个词汇的情感分值相加; 序列 4 表示程度副词与情感词汇有序且连续分布则进行分值的乘法运算, 比如在实际句子中出现“非常开心”则该单元作为分值为 2 的特征进行处理.

表 1 语义规则描述

序列号	符号描述	数学描述	例证
1	PS, PS	1+1	我是又开心又激动
2	PS, NS	1-1	内心有失落亦有欣慰
3	NS, NS	-1-1	真是伤心郁闷
4	Adv, S	2*S	非常开心
5	Neg, S	-1*S	过的不顺心

语义模型的规则匹配(rules)优先级: 本文中所采用的语义规则算法类似于数学计算, 数字越小代表优先集越高, 具体的语义规则优先级如表 2 所示, 特征词向量序列为[‘真的’, ‘开心’ ,‘快活’](用符号表示为[adv,ps,ps])时, 两个 ps 连续单元的优先级高于连续 adv, ps 单元, 则计算结果为 2*(1+1)=4.

表 2 语义规则优先级

优先级	序号	规则形式
1	1	PS, PS
	2	PS, NS
	3	NS, NS
2	4	Adv, S
	5	Neg, S

2.3 匹配流程

每次规则匹配结果作为新的情感分析基本单元, 也就是 S, 按照规则匹配的优先级进行迭代计算, 直到情感向量空间被完全匹配为止. 情感值的计算流程如图 1 所示. 针对基于规则匹配的方法对于反问句以及疑问句的判断难以做到精准识别的问题, 本文采取的策略是将问号和疑问词归并到否定词中, 以尽可能地减少误判漏判. 由于在微博签到数据中有很大一部分属于微博旅行话题范畴, 即统一带有#带着微博去旅行#这一语句段, 因此加入 word 分词词典, 在规则匹配中作为一次积极词汇进行计算. 图 1 给出了情感分析流程图.

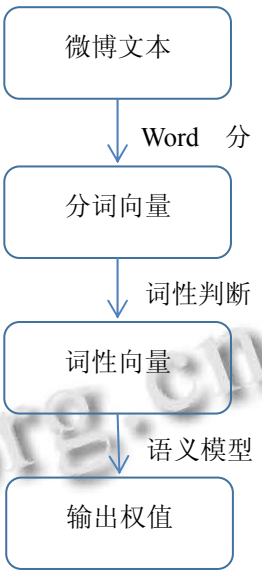

图 1 情感分析流程图

2.4 情感分析实验与结果分析

本文利用 Python 爬虫从新浪微博抓取了武汉范围内带有空间坐标信息的微博文本(有时间信息吗), 信息总量超过 2000 万条. 利用上文介绍的规则模板进行语义匹配, 计算得出了每条微博的情感值.

与基于监督学习的方法不同, 本文提到的方法得出的情感值并非分布于预先划定好的非连续空间, 每个句子的分析结果只有积极、消极、中立三种情况, 而是分布于连续的值空间, 即每个句子有一个具体的分值, 同时由于存在一些句子虽然由大量的积极词汇组成, 如连续带有多个大笑表情的句子, 虽然输出权值很高, 但并不一定代表其句子情感程度与普通带有积极成分的句子在真实世界存在如此巨大的差异, 因此, 本文将最终的输出权值的分布空间调整到-10~10, 其中大于 10 的句子权值统一记为 10, -10 的句子权值统一记为-10.

在计算准确率方面, 本文通过人工事先标注好 200 条微博文本样本, 用上文介绍的计算方法计算其情感值, 分别按照如表 3 和表 4 所示的阈值区间划分情感极性, 得出计算结果, 如表 5 所示, 经人工统计验证, 当按积极、消极、中立三个区间分割结果时, 准确率为 64.5%, 而按积极、消极两个区间分割结果时, 准确率则达到了 80%以上.

如微博文本“今天真是过得又开心又快活啊”, 依次经过文本分割、特征提取后变为[真是, 开心, 快活],

然后根据语义规则和优先级顺序计算为2*(1+1)=4, 因此这个句子的情感分值为 4.

表 3 情感阈值空间-1

区间	极性
<0.5	消极
-0.5~0,5	中立
>0.5	积极

表 4 情感阈值空间-2

区间	极性
<0	消极
>0	积极

表 5 情感分析结果

指标	结果
准确率(包含积极、消极、中立)	64.5%
准确率(只包含积极、消极)	80.5%

由此可见, 基于本文制定的语义规则模型的情感分类方法对正负情感具有较高的分类精度, 对于中立情感因为阈值选取以及人工标注的不确定性等原因使分类精度下降较大, 但总体仍呈现出较为满意的结果. 相对使用朴素贝叶斯的分类所产生的 74.54%的情感分类结果[9], 已经有了较大提升.

3 情感地图的制作

传统的情感分析通常仅仅用于挖掘纯粹的文本含义, 如对某一事件的社会舆论走向, 对某类产品或服务的评价等等[10]. 随着移动互联时代的到来, 越来越多的微博被附加以空间信息, 情感与时空的联系也有机会得以揭示.

幸福指数是人类内心依据自身状态所感受到的幸福感程度的判断, 幸福指数往往与人民生活水平, 对自身所处的环境、政策、基础设施的认可程度有关. 因此幸福指数作为一种新兴的社会指标正成为国家调整区域政策, 增强社会建设的一个新的依据. 传统的衡量城市幸福指数的方法几乎完全依赖于社会调查数据[11], 通常以主客观指标相结合的手段, 客观数据如 GDP 等, 主观数据主要来源于社会调查问卷. 因此传统的幸福指数评价体系需要耗费较大的人力物力, 而微博的供应员是人民大众, 也就是微博是取之于民的, 因而微博文本的情感极性直接反应了人民的情感状态, 并且考虑到微博的用户量之大、覆盖面之广, 这样以微博博文的情感极性值作为评判人民的幸福指数的主观指标是合理的.

情感地图是通过对地图的颜色渲染来展示不同区域人民的情感状态或对于某一事件的感情倾向. 幸福指数通常以数字列表的形式展现给人民群众, 本文提供的情感地图制作方法使情感数值鲜活地展现于地图之上, 不仅形象直观, 而且能够方便地观察不同区域人民情感极性的差异.

3.1 基于空间插值的情感地图

空间插值常用于将离散点的测量数据转换为连续的数据曲面, 以便与其它空间现象的分布模式进行比较. 经过情感计算的微博博文在空间上的分布是离散的, 通过空间插值推算出整个区域中情感值的连续变化, 在地图上以连续的颜色变化来展现数值上的差异, 由此得出的专题地图便是最终的情感地图. 图 2 给出了武汉市微博数据点分布地图, 其中获取的微博数据点超过 2000 万个, 图 3 给出了武汉市的情感色彩分布图, 其中, 欢乐谷、光谷和黄鹤楼都标注了最高的鲜艳红色, 代表这些旅游景点观光的人群具有较好的幸福感觉.

通过地图目视判读观察, 如图 3 所示, 武汉地区各大商圈以及旅游景点范围内的情感指数明显高于其周边范围, 实际的区域均值对照表如表 6 所示, 根据人类的认知经验, 购物和旅游是一种重要的休闲活动, 能调动人的积极情绪, 这也初步验证了情感地图表达信息的准确性.

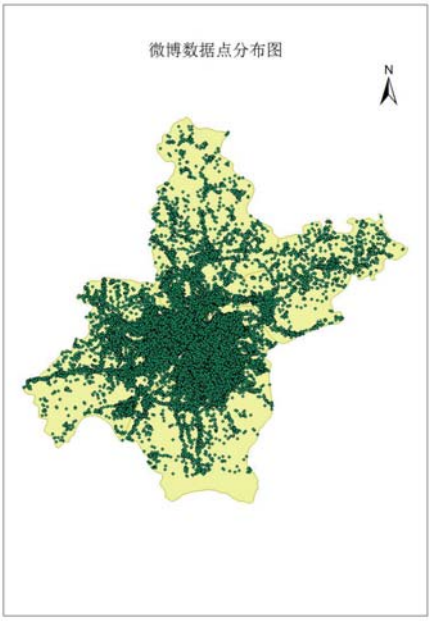

图 2 武汉市微博数据点分布地图

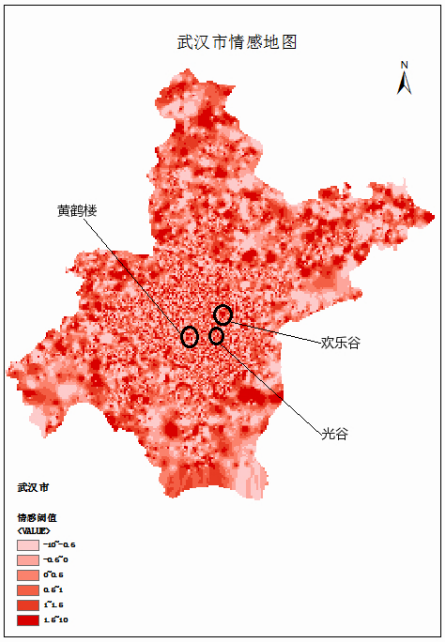

图 3 武汉市情感地图

表 6 代表区域情感分值对照表

地区	区域内情感均值	周边情感均值
黄鹤楼景区	0.51	0.34
欢乐谷乐园	0.56	0.38
光谷广场商圈	0.47	0.32

4 结语

本文针对微博文本的组成形式和特征, 使用附加有流行网络词汇和微博表情的情感词库, 采用自定义语义匹配规则的非监督学习方法对武汉市行政区内的微博进行了情感极性的判断, 然后利用地理信息系统(Geographic Information System, GIS)中的空间插值方法对情感微博进行了可视化的展示, 制作出了武汉市的情感地图, 将武汉市内的人民情感状态直观地绘制在地图之上, 同时作为一种新的评价公民幸福指数的手段, 具有易于获取、实时动态和客观真实等特点, 可以成为舆情监测、城市幸福指数评价的一种辅助手段, 从而为建立绿色宜居城市、城市规划发展发挥一定的辅助决策作用.

从情感分析的结果可以看出, 本文提出的方法对于中立情感识别的准确度仍需进一步加强, 在地图可视化中的重点区域可辨识性还较差, 对分析结果的验证还需要更科学的手段, 未来作者将针对这些问题进行进一步探索和改进. 对于中立情感准确度的提升, 有赖于良好的专家系统的评判, 并再此基础上通过交叉验证调整阈值的划分, 这一点将在后续的实验中进行实现并验证. 在地图可视化中的重点区域可辨识性还较差的问题, 将会在后续实验中对武汉市按区县、街区/商业区等进行区域的划分, 然后根据不同区域的情感均值赋予不同的等级色彩来提高情感地图的可辨识性.

参考文献

1 谢丽星,周明,孙茂松.基于层次结构的多策略中文微博情感分析和特征抽取.中文信息学报,2012,26(1):73–83.
2 崔连超.互联网评论文本情感分析研究.山东大学,2015.
3 孙艳,周学广,付伟.基于主题情感混合模型的无监督文本情感分析.北京大学学报:自然科学版,2013,49(1):102–108.
4 Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text. Proc. of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics. 2008. 1073–1080.
5 周胜臣,瞿文婷,石英子,等.中文微博情感分析研究综述.计算机应用与软件,2013,30(3):161–164.
6 Maas AL, Daly RE, Pham PT, et al. Learning word vectors for sentiment analysis. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. 2011. 142–150.
7 Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data. Proc. of the workshop on languages in social media. Association for Computational Linguistics. 2011. 30–38.
8 Hu X, Tang J, Gao H, et al. Unsupervised sentiment analysis with emotional signals. Proc. of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee. 2013. 607–618.
9 林江豪,阳爱民,周咏梅,等.一种基于朴素贝叶斯的微博情感分类.计算机工程与科学,2012,34(9):160–165.
10 Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2008, 2(1-2): 1–135.
11 Mitchell L, Frank MR, Harris KD, et al. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PloS one, 2013, 8(5): e64417.