中国科学: 信息科学 2014 年 第 44 卷 第 7 期: 825–835
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
双清论坛大数据专题 · 评述

公众文本之情感词典研究进展 ‡

饶洋辉[①*], 李青[①②], 刘文印[②], 李晶晶[①]

① 香港城市大学计算机科学系, 香港 999077
② 香港城市大学多媒体软件工程研究中心, 香港 999077
* 通信作者. E-mail: raoroland@gmail.com
‡ 特约编辑: 柴天佑, 张军, 王成红

收稿日期: 2013–12–13; 接受日期: 2014–01–21
香港城市大学策略研究基金项目 (批准号: 7002770) 资助

摘要 针对海量的社交媒体数据进行情感分析, 能够实时检测与跟踪公众对社会事件、政治活动、公司战略、重大决策等方面的观点, 同时了解用户在其评论、博客、微博等文本中的情感倾向. 本文首先论述了从作者角度出发、针对文本中的主观词等构建的文本情感词典研究现状, 包括基于辞典的和基于语料库的生成方法及其典型应用. 然后阐述了从读者角度出发、通过读者对文档的情感反馈而构建的公众情感词典研究进展, 包括其数据来源, 以及词层和主题层的公众情感词典生成方法和模型.

关键词 社交媒体 情感词典 情感分析 公众情感检测 大数据 主题模型

1 引言

Web 2.0 应用自 2004 年发展以来, 从论坛、博客、社交网络等在线社交媒体中产生了海量的用户生成内容. 这些数据量大而复杂的应用中产生的数据集 (包括社交媒体数据) 及其分析技术分别被称为 "大数据" 和 "大数据分析"[1]. Web 2.0 应用不仅能高效地为不同商业提供其用户的实时反馈与观点, 而且能够使公众在社交媒体中表达其社会及政治情感 [1]. 社交媒体数据有助于理解评论者和公众关于社会事件、政治活动、公司战略、市场营销和产品偏好等方面的观点 [2]. 评论者观点的社交媒体分析主要基于文本分析、情感分析等技术 [2], 而情感分析又取决于对文本中的情感词及其极性 (polarity, 如正面和负面) 的识别 [3]. 因此, 情感词典在文本情感分析中具有重要作用. 此外, 大部分公众情感检测方法也基于情感词典, 因为机器学习方法通常需要大量及多样化的人工标注的社会观点数据 [4]. 由此可见, 情感词典对于文本情感分析和公众情感检测均有较大的应用价值.

本文从生成方法、典型应用和数据来源等角度依次对文本情感词典和公众情感词典进行评述, 其主要结构如下: 第 2 章论述当前开发的文本情感词典, 重点介绍基于辞典的和基于语料库的生成方法, 以及文本情感词典的应用; 第 3 章阐述公众情感词典, 包括数据来源, 以及词层、主题层的公众情感词典构建方法与模型; 第 4 章为全文总结.

引用格式: 饶洋辉, 李青, 刘文印, 等. 公众文本之情感词典研究进展. 中国科学: 信息科学, 2014, 44: 825–835, doi: 10.1360/N112013-00167

表 1　GI 基本电子表单的类别示例

Table 1　Examples of general inquirer basic spreadsheet categories

Category	Number of words	Examples
Positiv (positive)	1,915	Ability, accept
Negativ (negative)	2,291	Angry, defame
Strong (strong)	1,902	Abundant, busy
Weak (weak)	755	Avoid, cease
Active (active)	2,045	Achieve, compute
Passive (passive)	911	Ail, omit
Virtue (virtue)	719	Adroit, cute
Vice (vice)	685	Bad, deceit
Ovrst (overstated)	696	Bulk, entire
Undrst (understated)	319	Barely, cursory
Pleasur (pleasure)	168	Admire, enjoy
Pain (pain)	254	Agony, dread

2　文本情感词典

文本情感词典是从作者角度出发、针对文本中的主观词等构建的情感词典. 文本情感词典的生成主要包括基于辞典的和基于语料库的方法. 文本情感词典通常用于情感自动标注、极性分类等研究与应用中.

2.1　基于辞典的生成方法

目前, 常用的辞典有英文的 WordNet[5]1)和 General Inquirer (GI)[6]2), 中文的知网 (HowNet)3)等. WordNet 辞典是英文的词汇语义网络系统, 它将英文的名词、动词、形容词和副词组织为同义词集合 (synsets), 每一个集合表示一个基本的词汇概念, 并在这些词汇概念间建立同义关系、反义关系、上下位关系、部分关系和完全关系等多种词汇语义关系 [7]. WordNet 辞典包含 11 万 7 千个同义词集合及其概念关系, 该辞典已被应用于词义消歧、计算语言学、文本分析、双语及多国语机器翻译、信息检索、情感分析等一系列领域中. GI 是一个包含词性、语法、语义、语用信息的英文辞典 [6], 在计算应用中常用的是其电子表单 (spreadsheet). GI 的基本电子表单 (general inquirer basic spreadsheet) 包括 11,788 行及 184 列. 其中, 第 1 列展示每个英文词或词义, 第 2 列指明该词是否被 Harvard 或 Lasswell 词典收录, 其余 182 列为各词条所属的类别. 词条数较多的类别及相关信息如表 1 所示. 知网辞典是一个以汉语和英语的词语所代表的概念为描述对象, 以揭示概念之间及概念所具有的属性之间的关系为基本内容的常识知识库. 与 WordNet 辞典类似, 知网描述了同义关系、反义关系、对义关系、上下位关系、部件 – 整体关系、事件 – 角色关系、相关关系等.

1) http://wordnet.princeton.edu/.
2) http://www.wjh.harvard.edu/∼inquirer/.
3) http://www.keenage.com/.

基于 WordNet 辞典生成的情感词典有 WordNet-Affect[8] 和 SentiWordNet[9~11] 等. WordNet-Affect 情感词典是对 WordNet 辞典中的同义词集合进行情感标注, 其过程如下: 首先人工挑选了 1903 个 "核心" 情感词, 包括名词 (539 个)、形容词 (517 个)、动词 (238 个) 和副词 (15 个) 等, 然后利用 WordNet 辞典中词之间的同义关系扩充情感词典, 最后共标注了 2874 个同义词集合与 4787 个词. 与之相关的工作包括: Hu 和 Liu[12], 以及 Kim 和 Hovy[13] 所提出的各词的同义词 (反义词) 与原词具有相同 (相反) 的极性, 这些同义词和反义词均通过 WordNet 辞典进行查询. 据此, 目标词的极性由其邻近的正面词或负面词的个数决定. Godbole 等 [3] 提出了一种大规模的新闻与博客情感分析系统, 并对前述方法进行了改进, 其思路如下: 首先基于 WordNet 辞典查询各词的同义词 (反义词), 然后不断查询其同义词 (反义词) 的同义词及反义词, 并使极性的大小随查询的层级而递减. 此外, Kamps 等 [14] 以词为节点、WordNet 辞典词之间的同义关系为边生成图, 然后通过计算目标词与词 "good" 和 "bad" 之间的最短距离来衡量其极性大小. SentiWordNet 情感词典主要包括 SentiWordNet 1.0[9]、SentiWordNet 2.0[10] 和 SentiWordNet 3.0[11]3 种公开版本, 其采用半监督的三元分类器和随机游走等对 WordNet 辞典中的同义词集合进行情感标注, 分别计算正面、负面和中性 3 种分值并得到最终的极性大小.

基于知网 (HowNet) 辞典构建的情感词典如 "Chinese/English Vocabulary for Sentiment Analysis" (VSA)[4] . VSA 情感词典包含中文和英文 2 种语言, 每种语言均列出了 6 种词列表, 分别为: 正面情感词语、负面情感词语、正面评价词语、负面评价词语、程度级别词语和主张词语. 中文的正面情感词语如 "爱"、"表扬"、"感恩"、"高兴" 等; 负面情感词语如 "哀伤"、"悲切"、"动怒" 等; 正面评价词语如 "安全"、"百里挑一"、"便捷" 等; 负面评价词语如 "昂贵"、"笨重"、"不灵敏" 等; 程度级别词语如 "非常"、"极其"、"万分" 等; 主张词语如 "察觉"、"认为"、"相信" 等. 在此基础上, 增强版的情感词典中又新增了 6 种词语类别 [5], 即正面属性、负面属性、正面实体、负面实体、正面事件和负面事件.

2.2 基于语料库的生成方法

除了上述基于 WordNet, GI, HowNet 等辞典的生成方式, 还有基于语料库生成的文本情感词典. Hatzivassiloglou 和 McKeown[15] 通过对 "1987 Wall Street Journal" 语料库 [6]中 2 千 1 百万个词的分析, 得出了如下规则: 以 "and" 分隔的形容词具有相同的极性, 而以 "but" 分隔的则为相反的极性. 据此, 从一个较小的种子列表出发, 可以采用该规则生成及扩充文本情感词典. Turney[16] 首先定义了一个正面词 (如词 "excellent") 和负面词 (如词 "poor") 的种子集, 然后利用 AltaVista 搜索引擎收录的 3 亿 5 千万个英文网页作为语料库, 分别计算目标词与正、负面种子词的点互信息 (point-wise mutual information, PMI), 并将其差值作为目标词的极性. Banea 等 [17] 基于一个较小的种子集 (60 个词)、在线词典, 以及较小规模的语料库 (50 万个词) 生成罗马尼亚语的主观词词典 (subjectivity lexicon). 由匹兹堡大学、康乃尔大学和犹他大学研究人员开发的 OpinionFinder 系统 [7]也包含一个主观词词典 [18], 将其整理后的词条如表 2 所示. 该情感词典源于 MPQA 观点语料库 (multi-perspective question answering opinion corpus)[19,20]. MPQA 是一种对多种来源的新闻文档人工标注其观点、情感等状态的语料库, 它同时提供了主观感觉注释文档 (subjectivity sense annotations).

最近, Peng 和 Park[4] 提出了 WordNet 辞典和大规模语料库相结合的情感词典生成方法, 实验结

4) http://www.keenage.com/html/e_bulletin_2007.htm.
5) http://www.keenage.com/html/General_statistics_of_sentiment_words_and_expressions.htm.
6) http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2000T43.
7) http://mpqa.cs.pitt.edu/opinionfinder/.

表 2　OpinionFinder/MPQA 的主观词词典示例

Table 2　Examples of OpinionFinder/MPQA subjectivity lexicon

Word	Part-of-speech	Strength	Polarity
Abandoned	Adj (adjective)	Weaksubj (weak subjective)	Negative
Abandonment	Noun (noun)	Weaksubj (weak subjective)	Negative
Abandon	Verb (verb)	Weaksubj (weak subjective)	Negative
Abase	Verb (verb)	Strongsubj (strong subjective)	Negative
Zest	Noun (noun)	Strongsubj (strong subjective)	Positive

果显示 2 种方式相结合的效果要优于其中任何一种方式. Xu 等 [21] 也提出了一种基于辞典与语料库生成中文情感词的方法, 其情感维度包括高兴、愤怒、悲伤、恐惧和震惊. 采用类似方法生成的情感词典如 LIWC (linguistic inquiry and word count) 词典 [22], 该词典包含正面情感、负面情感、愤怒、压抑等 64 种情感、行为和心理维度. Thelwall 等 [23] 则基于初始辞典、机器学习方法、拼写校正法等开发了 SentiStrength 工具 8), 该工具能够输出单个及多个词的情感得分.

2.3　文本情感词典的应用

文本情感词典的常见应用是对观点文档 (如观点评述型的微博、评论者撰写的评论等) 进行情感自动标注与极性分类. 例如, Lansdall-Welfare 等 [24] 将 WordNet-Affect 词典用于对 4 亿 8 千多万条英文微博进行情感标注, 其自动标注的维度包含高兴、恐惧、愤怒和悲伤 4 种情绪, 结果显示自动标注的情感与人工的判读相吻合. O'Connor 等 [25] 基于 OpinionFinder/MPQA 主观词词典, 对 10 亿条 Twitter 微博文本进行正面和负面情感的极性分类, 其与问卷调查结果的相关度达到了 80%. Miller 等 [26] 通过 General Inquirer、SentiWordNet 词典, 以及表情符号分析对社交网络的情感流动进行研究, 其社交网络图含有将近 8 百万个节点和 1 千 5 百万条边. Golder 和 Macy[27] 基于 LIWC 词典对 5 亿多条 Twitter 微博文本进行极性分类, 据此分析其 2 百多万微博用户的情绪变化. Kucuktunc 等 [28] 采用 SentiStrength 工具计算 Yahoo! 问答中 3 千 4 百万条提问、1 亿 3 千 2 百万条回答、4 亿 1 千 2 百万个句子的情感得分, 并对大数据环境下情绪与文本长度、标点符号、人口特征等的相互关系进行了详尽的分析.

文本情感词典在实际应用中的主要优势为简便快速, 因而能够对大数据进行实时情感分析, 如文本情感词典在社会渠道分析平台 LCI (惠普实验室的原型系统) 等工业领域中的应用 [29]. 然而, 由于文本情感词典通常是作为通用领域的参考源, 其在不同语言和领域中的应用效果可能有较大差别. 总体而言, 文本情感词典的应用所面临的主要问题如下 [30]:

(1) 同一个词在不同语境、领域下具有不同的极性. 例如, 词 "很大" 在句子 "这台冰箱耗电很大" 中是负面情感, 而在句子 "这间旅馆的房间很大" 中是正面情感.

(2) 同一个词具有不同的含义. 例如, 单词 "suck" 具有 "讨厌" 和 "吸" 等不同的含义. 句子 "This camera sucks" 指的是照相机不好 (负面评价); 句子 "This vacuum cleaner really sucks" 指的则是吸尘器很有效 (正面评价).

(3) 含有主观词的文本可能没有表达情感. 通常, 提问句中会含有主观词, 如句子 "请问这部电影好看吗?", 但它们并未包含正面或负面的情感.

8) http://sentistrength.wlv.ac.uk/.

表 3 含有情感投票功能的中文网站示例 a)

Table 3 The websites containing an emotion rating component in China

Website	Channel	Emotion labels
People.com.cn	Opinion	Shocked, anger, sadness, touched, pleased, happiness, boredom, amusement
Sina.com.cn	Society	Touched, empathy, boredom, anger, amusement, sadness, surprise, warmness
Chinanews.com	Culture	Touched, empathy, boredom, anger, amusement, sadness, pleased, careless
Huanqiu.com	Technology	Shocked, anger, sadness, touched, pleased, happiness, boredom, amusement
QQ.com	Entertainment	Pleased, touched, empathy, anger, amusement, sadness
Sohu.com	Education	Pleased, anger, sadness, cool, crazy

a) 新浪网社会频道的情感数已由之前的 8 个减至 6 个, 分别为: 感动 (touched)、震惊 (shocked)、搞笑 (amusement)、难过 (sadness)、新奇 (surprise)、愤怒 (anger).

(4) 不含主观词的文本可能表达情感. 如 “我看了这部电影, 中途睡了一个小时” 等客观反映事实的文本也可能隐含了用户的情感.

3 公众情感词典

公众情感词典是从读者角度出发、通过读者对文档的情感反馈而构建的情感词典. 公众情感词典通常采用有监督的方法进行自动构建, 主要分为词层和主题层两类. 基于句子层或文档层的情感标注文档, 可以得到每个词或主题的公众情感分布情况, 有助于检测及跟踪特定实体 (如人物、机构、地点、品牌等) 和事件所引发的公众情感.

3.1 数据来源

目前, 公众可以对新闻文档进行情感投票的中文网站如表 3 所示. 以环球网的科技频道为例, 网站每发布一篇科技类的新闻文档, 均在文档后面附上了 “震惊”、“愤怒”、“悲伤” 等 8 种情感标签. 每位用户看完一篇新闻后, 都可以选择某种情感标签进行投票以表达其对这篇新闻的 “读后感”. 这些包含公众情感投票数的文档即为公众情感词典的数据来源.

相比上述多样的中文数据来源, 当前含有类似情感投票功能的英文网站较少. 英文版的北美新浪网站在其社会频道 [9]虽然也提供了 8 种情感标签, 但参与投票的用户较少. 2007 年, 第四届语义评价国际研讨会 (SemEval) 提供了一种英文数据集 [10]. 该数据集包含愤怒 (anger)、厌恶 (disgust)、恐惧 (fear)、高兴 (joy)、难过 (sad) 和惊讶 (surprise)6 种情感, 对于每篇英文新闻标题 (源自纽约时报、CNN、BBC 新闻和 Google 新闻), 均由人工依据上述情感维度进行投票, 其值的大小从 0 到 100. 然而, 该数据集仅包括 1246 篇总投票数大于 0 的新闻标题. 由此可见, 当前英文数据来源的规模仍较小.

3.2 词层公众情感词典研究

词层公众情感词典可用于检测及跟踪人物、机构、地点、品牌等特定实体所引发的公众情感. 此

9) http://english.sina.com/news/china/society.html.

10) http://www.cse.unt.edu/∼rada/affectivetext/.

类研究是将包含情感投票数据的文档划分为词, 然后生成词层的公众情感词典. 以下详述 3 种词层公众情感词典的生成方法:

SWAT 系统 [31] 基于 SemEval 公开的英文数据集, 采用一种有监督的方法生成词层的公众情感词典. 记词 w_j在情感 e_k上的得分为 Score(e_k, w_j), 所有包含词 w_j的新闻标题为 $\mathbb{H}$, 公众针对 $\mathbb{H}$ 在情感 e_k上的投票数为 Score$(e_k, \mathbb{H})$, 则:

$$\text{Score}(e_k, w_j) = \sum_{\mathbb{H}: w_j \in \mathbb{H}} \text{Score}(e_k, \mathbb{H}). \tag{1}$$

由此可见, SWAT 系统采用直观的加总方式生成词层的公众情感词典, 该词典的项即为Score(e_k, w_j).

Emotion-Term (简称 ET) 方法 [32,33] 依据朴素 Bayes 分类器的思想来生成词层的公众情感词典, 该词典的项, 即给定情感 e_k, 出现词 w_j的条件概率估计式子如下:

$$P(w_j|e_k) = \frac{|(w_j, e_k)|}{\sum_{w \in \mathbb{W}} |(w, e_k)|}, \tag{2}$$

其中, $\mathbb{W}$是所有词的集合, $|w_j, e_k|$是所有文档中词 w_j和情感 e_k共现的次数, 其计算式子如下:

$$|(w_j, e_k)| = S + \sum_{d \in \mathbb{D}} c_{d,j} \cdot r_{d,k}, \tag{3}$$

上式中, $\mathbb{D}$是所有文档的集合, S是一个较小的平滑常数 (如常数 1), $c_{d,j}$表示文档 d中词 w_j出现的频次, $r_{d,k}$表示公众针对文档 d在情感 e_k上的投票数.

ET 方法采用的新浪网社会频道的中文数据集, 该数据集包含 2858 篇新闻文档, 以及公众在 8 种情感标签上的 659174 次投票.

Word-Emotion (简称 WE) 方法 [34] 基于极大似然估计等生成词层的公众情感词典. 该词典的项, 即给定词 w_j, 出现情感 e_k的条件概率计算式子如下:

$$P(e_k|w_j) = \frac{\sum_{d \in \mathbb{D}} \varepsilon_d \sigma_{d,j} r_{d,k}}{\sum_{e \in \mathbb{E}} \sum_{d \in \mathbb{D}} \varepsilon_d \sigma_{d,j} r_{d,e}}, \tag{4}$$

其中, $\mathbb{E}$是所有情感 (标签) 的集合, $\sigma_{d,j}$表示文档 d中词 w_j出现的相对频次, $r_{d,k}$表示公众针对文档 d在情感 e_k上的投票数, $r_{d,e}$表示公众针对文档 d在情感 e上的投票数, ε_d为文档 d出现的先验概率, 其估计式子如下:

$$\varepsilon_d = \sum_{e \in \mathbb{E}} (r_{d,e} / \sum_{d' \in \mathbb{D}} r_{d',e}). \tag{5}$$

WE 方法也采用了新浪网社会频道的中文数据集, 该数据集包含 40,897 篇新闻文档, 以及公众在 8 种情感标签上的 2083818 次投票. 由于词性信息 (如名词、动词、形容词、副词等) 可用于词义的消歧 [35], 将词性标注用于改进上述情感词典的生成方法 [36] 能够发现词形相同而词性不同的词可能引发的不同情感, 且有助于对人物、机构、地点、品牌等特定实体的判别.

3.3 主题层公众情感词典研究

主题层公众情感词典可用于检测及跟踪特定事件所引发的公众情感. 此类研究首先将包含情感投票数据的文档划分为主题, 然后生成主题层的公众情感词典. 以下详述两种主题层公众情感词典的生成模型:

表 4 公众情感的代表主题示例

Table 4 The representative topics of social emotions

Emotion label	Topic ID	Top words in each topic
Sadness	1	Suicide, accident, incident, salvage, corpse, stress
Surprise	2	Award, kilogram, bonuses, snake, weight, cats
Anger	3	Suspect, judgments, public security, authority, rape, intentionally
Empathy	4	Work, sister, brother, disease, illness, serious
Amusement	6	Even, divorce, website, young lady, woman, actually
Boredom	8	Networks, college students, net, relationships, women, media
Warmness	11	Bride, wedding, groom, sisters, happy, new couple
Empathy	13	Hometown, head, black bear, regulations, euthanasia, surrender
Surprise	15	Lotteries, study, award, win, phase,record
Touched	16	Save, touched, insist, transplant, story, thanks

Emotion-Topic Model (简称 ETM)[32,33] 是一种对文档潜在主题和情感投票进行联合建模的模型, 其生成过程如下:

(1) 从文档相关的公众情感投票分布中产生一次情感投票;

(2) 由上述情感投票得到的多项式分布中生成一个潜在主题;

(3) 由上述潜在主题得到的多项式分布中生成一个词.

ETM 采用了与上节中 ET 方法所使用的相同数据集, 据此得到每种公众情感的代表主题, 如表 4 所示. 其中, "同情" 的代表主题为 4 和 13; "新奇" 的代表主题为 2 和 15.

Emotion LDA (简称 ELDA) 模型 [34] 同样是对文档潜在主题和情感投票进行联合建模. 基于极大似然估计等方法, ELDA 能够生成主题层的公众情感词典. 该词典的项, 即给定主题 z_m, 出现情感 e_k的条件概率计算式子如下:

$$P(e_k|z_m) = \frac{\sum_{d\in\mathbb{D}} \varepsilon_d P(z_m|d) r_{d,k}}{\sum_{e\in\mathbb{E}} \sum_{d\in\mathbb{D}} \varepsilon_d P(z_m|d) r_{d,e}}, \tag{6}$$

其中, ε_d, $r_{d,k}$和 $r_{d,e}$ 的含义同式 (4), $P(z_m|d)$为给定文档 d, 出现主题 z_m的条件概率, 其估计式子如下:

$$P(z_m|d) = \frac{n^d_{z_m} + \alpha}{\sum_{z\in\mathbb{Z}}(n^d_z + \alpha)}, \tag{7}$$

上式中, $\mathbb{Z}$是所有主题的集合, $n^d_{z_m}$表示文档 d中的词被指定为主题 z_m的个数, α为超参数.

ELDA 采用了与上节中 WE 方法所使用的相同数据集, 据此得到每种主题所引发的 (代表性) 公众情感, 如表 5 所示. 其中, 主题 375 为 "混合型" 主题, 它引发了公众的多种不同极性的情感; "公众情感分布" 中括号里的数值为给定主题 z_m, 出现各种情感的条件概率, 其由式 (6) 计算得知; "主题的代表词" 中括号里的数值为给定主题 z_m, 出现词 w_j的条件概率, 其估计式子如下:

$$P(w_j|z_m) = \frac{n^{z_m}_{w_j} + \beta}{\sum_{w\in\mathbb{W}}(n^{z_m}_w + \beta)}, \tag{8}$$

上式中, $n^{z_m}_{w_j}$表示词 w_j 被指定为主题 z_m的次数, β为超参数.

主题层的公众情感词典不仅比词层的更易理解 [37,38], 且能更好地反映文档的多主题特性. 例如,

表 5　主题所引发的公众情感示例

Table 5　Examples of the topics evoking social emotions

Topic ID	Top words in each topic	Distribution of social emotions
13	Help(0.004), Touched (0.004), hope (0.003)	Touched (0.47), sadness (0.11), empathy (0.11)
126	Dog (0.015), animal (0.015), breeding (0.010)	Surprise (0.21), amusement (0.18), anger (0.15)
212	Death (0.012), rescue (0.009), die (0.009)	Sadness (0.27), anger (0.26), empathy (0.14)
219	Hospital (0.011), doctor (0.010), cure (0.010)	Touched (0.20), empathy (0.18), sadness (0.16)
375	University (0.020), student (0.015), school(0.012)	Touched (0.19), anger (0.18), amusement (0.15)

文档中可能有一些主题外 (off-topic) 的段落, 它们传达的情感信息与本主题并不相关, 词层的公众情感词典难以对此进行判别 [35]. 然而, 由于主题层的公众情感词典生成算法复杂度比词层的更高, 如何在大数据中实时生成及应用主题层的公众情感词典是其面临的主要挑战之一. 近年来, 分布式算法已被用于对百万级的文档进行潜在主题的建模 [39], 这为大数据中主题层的公众情感词典研究提供了新途径.

4　结语

情感词典是情感分析的基石, 本文将其分为文本情感词典和公众情感词典两大部分进行评述. 对于文本情感词典部分, 首先阐述了基于辞典的和基于语料库的文本情感词典生成方法, 然后论述了文本情感词典在情感自动标注和极性分类中的应用. 除了基于文本情感词典的极性分类方法, 其它的还有将分类器直接用于极性分类的研究. 例如, Twitter 公司的研究人员 Lin 和 Kolcz [40] 在基于 Hadoop 的分析平台中, 通过逻辑回归分类器对海量的微博数据进行极性分类. 实验采用了 3 种规模大小的训练集, 分别为 1 百万、1 千万和 1 亿条微博, 测试集大小为 1 百万条微博. 对于公众情感词典部分, 首先重点介绍了中、英文的数据来源, 然后详述了词层、主题层的公众情感词典研究. 相关的研究领域除了对公众情感词典的生成方法和模型的探讨, 还有基于文本情感词典 [41,42], 以及分类器 [43,44] 的公众情感挖掘研究.

随着数据来源的多样化, 以及自然语言文本解析技术的发展, 情感分析将从基于词的研究向概念、背景 [45]、主题等维度的深入. 但当前概念、背景、主题等维度的情感分析模型复杂度较高, 因而仍需借鉴分布式等算法提高其效率以适应持续增大的数据量.

参考文献

1 Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: from big data to big impact. MIS Quart, 2012, 36: 1–24

2 Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retrieval, 2008, 2: 1–135

3 Godbole N, Srinivasaiah M, Skiena S. Large-scale sentiment analysis for news and blogs. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, 2007

4 Peng W, Park D H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2011

5 Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998

6 Philip J S, Dunphry D C, Smith M S, et al. The General Inquirer: A Computer Approach to Content Analysis. Cambridge: MIT Press, 1966

7 Miller G A, Beckwith R, Fellbaum C, et al. WordNet: an on-line lexical database. Int J Lexicography, 1990, 3: 235–244

8 Strapparava C, Valitutti A. Wordnet-affect: an affective extension of wordnet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004. 1083–1086

9 Esuli A, Sebastiani F. Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006. 417–422

10 Esuli A. Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms, and Applications. PhD Thesis. Pisa: University of Pisa, 2008

11 Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th Conference on Language Resources and Evaluation, 2010. 2200–2204

12 Hu M Q, Liu B. Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2004. 168–177

13 Kim S M, Hovy E. Determining the sentiment of opinions. In: Proceedings of the Coling Conference, Stroudsburg, 2004

14 Kamps J, Marx M, Mokken R J, et al. Using WordNet to measure semantic orientation of adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004. 1115–1118

15 Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives. In: Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, Stroudsburg, 1997. 174–181

16 Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Stroudsburg, 2002. 417–424

17 Banea C, Mihalcea R, Wiebe J. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008

18 Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase–level sentiment analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Stroudsburg, 2005. 347–354

19 Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Lang Resour Eval, 2005, 39: 165–210

20 Wilson T. Fine-Grained Subjectivity Analysis. PhD Thesis. Pittsburgh: University of Pittsburgh, 2008

21 Xu G, Meng X F, Wang H F. Build Chinese emotion lexicons using a graph–based algorithm and multiple resources. In: Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, 2010. 1209–1217

22 Pennebaker J W, Francis M E, Booth R J. Linguistic Inquiry and Word Count (LIWC): LIWC2001. Mahwah: Erlbaum Publishers, 2001

23 Thelwall M, Buckley K, Paltoglou G, et al. Sentiment in short strength detection informal text. J Am Soc Inf Sci Tec, 2010, 61: 2544–2558

24 Lansdall-Welfare T, Lampos V, Cristianini N. Effects of the recession on public mood in the UK. In: Proceedings of Mining Social Network Dynamics session on Social Media Applications in News and Entertainment at World Wide Web, New York, 2012. 1221–1226

25 O'Connor B, Balasubramanyan R, Routledge B R, et al. From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010

26 Miller M, Sathi C, Wiesenthal D, et al. Sentiment flow through hyperlink networks. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, 2011

27 Golder S, Macy M. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science, 2011, 333: 1878–1881

28 Kucuktunc O, Cambazoglu B, Weber I, et al. A large-scale sentiment analysis for Yahoo! answers. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, New York, 2012. 633–642

29 Castellanos M, Dayal U, Hsu M, et al. LCI: a social channel analysis platform for live customer intelligence. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, New York, 2011. 1049–1058

30 Liu B. Sentiment Analysis and Opinion Mining. San Rafael: Morgan & Claypool Publishers, 2012. 1–167

31 Katz P, Singleton M, Wicentowski R. Swat–mp: the semeval–2007 systems for task 5 and task 14. In: Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, 2007. 308–313
32 Bao S H, Xu S L, Zhang L, et al. Mining social emotions from affective text. IEEE Trans Knowl Data Eng, 2012, 24: 1658–1670
33 Bao S H, Xu S L, Zhang L, et al. Joint emotion-topic modeling for social affective text mining. In: Proceedings of the 9th IEEE International Conference on Data Mining, Miami, 2009. 699–704
34 Rao Y H, Lei J S, Liu W Y, et al. Building emotional dictionary for sentiment analysis of online news. World Wide Web J, 2013. 1–20
35 Cambria E, Schuller B, Xia Y Q, et al. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst, 2013, 28: 15–21
36 Lei J S, Rao Y H, Li Q, et al. Towards building a social emotion detection system for online news. Future Gener Comp Sy, 2013
37 Quan C Q, Ren F J. An exploration of features for recognizing word emotion. In: Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, 2010. 922–930
38 Rao Y H, Li Q, Mao X D, et al. Sentiment topic models for social emotion mining. Inform Sciences, 2014, 266: 90–100
39 Newman D, Asuncion A, Smyth P, et al. Distributed algorithms for topic models. J Mach Learn Res, 2009, 10: 1801–1828
40 Lin J, Kolcz A. Large-scale machine learning at twitter. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2012. 793–804
41 Chaumartin F R. Upar7: a knowledge–based system for headline sentiment tagging. In: Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, 2007. 422–425
42 Strapparava C, Mihalcea, R. Semeval–2007 task 14: affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, 2007. 70–74
43 Lin K H Y, Yang C, Chen H H. Emotion classification of online news articles from the reader's perspective. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Washington, 2008. 220–226
44 Lin K H Y, Yang C, Chen H H. What emotions do news articles trigger in their readers? In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 2007. 733–734
45 Cambria E, Rajagopal D, Olsher D, et al. Big social data analysis. Big Data Comput, 2013. 401–414

Progress of generating sentiment lexicons from text in social media

RAO YangHui[1*], LI Qing[1,2], LIU WenYin[2] & LI JingJing[1]

1 *Department of Computer Science, City University of Hong Kong, Hong Kong* 999077*, China;*
2 *Multimedia software Engineering Research Centre, City University of Hong Kong, Hong Kong* 999077*, China*
*E-mail: raoroland@gmail.com

Abstract Sentiment analysis for big social media data could not only detect and track public opinions about social events, political movements, company strategies and decisions in real time, but also understand users emotions which are expressed through reviews, blogs, microblogs/tweets and other text. In this article, we first present the progress of generating sentiment lexicons in text with subjective words from the writer's perspective, including the thesaurus-based and corpora-based methods and typical applications. Then, the progress of generating social

emotion lexicons by the emotional responses of readers from the reader's perspective is reviewed, which includes the data sources, word-level and topic-level models.

Keywords social media, emotional dictionary, sentiment analysis, social emotion detection, big data, topic model

RAO YangHui was born in 1986. He received his Master's degree from Graduate University of Chinese Academy of Science at 2010. Currently, he is pursuing his Ph.D. degree in the department of Computer Science, City University of Hong Kong. His research interest covers sentiment analysis, natural language processing and text mining.

LI Qing was born in 1962. He received the B.E. degree from Hunan University, China, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, USA, all in computer science. He is now a professor at the City University of Hong Kong. His research interests include data mining, Web services, multimedia retrieval and management, and e-learning systems. Prof. Li serves as the chairman of the Hong Kong Web Society, a councilor of the Database Society of China Computer Federation, and a Steering Committee member of DASFAA, ICWL, WAIM, and International WISE Society. He is a Fellow of IET and a Senior Member of IEEE.

LIU WenYin was born in 1966. He has a B.E. and M.E. in computer science from Tsinghua University, Beijing and a DSc from the Technion, Israel Institute of Technology, Haifa. He is current a senior research fellow at the multimedia software engineering research center, City University of Hong Kong. He was an assistant professor at the City University of Hong Kong from 2002 to 2012 and a full time researcher at Microsoft Research China/Asia from 1999 to 2001. His research interests include anti-phishing, question answering, graphics recognition, and performance evaluation. In 2003, he was awarded the International Conference on Document Analysis and Recognition Outstanding Young Researcher Award by the International Association for Pattern Recognition (IAPR). He is a Fellow of IAPR and a Senior Member of IEEE.

LI JingJing was born in 1982. She received the Ph.D. degree from Hong Kong Polytechnic University in 2012. Currently, she is a lecturer in School of Computer Science at South China Normal University. Her research interests are mainly focus on evolutionary algorithm, energy efficient routing and object tracking for wireless sensor networks.