DOI:10.16180/j.cnki.issn1007-7820.2014.11.052

专题综述

文本情感分析研究综述

马 力[1]，宫玉龙[2]
(1. 西安邮电大学 数字艺术学院，陕西 西安 710121；2. 西安邮电大学 计算机学院，陕西 西安 710061)

摘 要 对文本情感分析的研究现状与进展进行总结。从情感分析的任务情感分类、情感检索、情感抽取 3 个方面详细介绍了相关研究和技术方法，重点阐述了基于语义的情感词典分类方法和基于机器学习的情感分类方法，并介绍了文本情感分析的评测，提出了未来的研究方向。

关键词 情感分析；情感分类；情感检索；情感抽取
中图分类号 TP393 文献标识码 A 文章编号 1007－7820(2014)11－180－05

Research on Text Sentiment Analysis

MA Li[1], GONG Yulong[2]
(1. School of Digital Arts, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract This paper reviews the research status and progress of text sentiment analysis. Three tasks of sentiment analysis are analyzed in detail: sentiment classification, sentiment retrieval, and sentiment extraction with focus on the sentiment classification methods based on sentiment dictionary and machine learning. The evaluation of sentiment analysis is introduced. Finally the trend of text sentiment analysis is concluded.

Keywords sentiment analysis; sentiment classification; sentiment retrieval; sentiment extraction

随着互联网的飞速发展，尤其是微博、微信等新社交网络的兴起，大量网络用户每天都会发布并传播高达上亿的信息。这些海量的文本信息中，有很大一部分是表达用户观点倾向和情感信息，如支持、反对、喜、怒、哀、乐等的文本信息。这些情感文本信息是非常宝贵的意见资源，包含着人们对社会各种现象的不同观点和立场，话题涉及政治、经济、军事、娱乐、生活等领域。例如一个人想买一件商品，通常先在网上查看该商品的相关评论。个人和组织越来越多地把网络上的情感观点信息用于决策，因此使用计算机技术自动对其分析处理，在话题跟踪发现、舆情跟踪、民意测验、定向广告投放，售后服务评价等领域有着广泛的应用前景，情感分析技术应运而生。

1 情感分析简述

文本情感分析(Sentiment Analysis)是指利用自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分析、处理和抽取的过程[1]。

目前，文本情感分析研究涵盖了包括自然语言处理、文本挖掘、信息检索、信息抽取、机器学习和本体学等多个领域，得到了许多学者以及研究机构的关注，近几年持续成为自然语言处理和文本挖掘领域研究的热点问题之一。情感分析任务按其分析的粒度可以分为篇章级，句子级，词或短语级；按其处理文本的类别可分为基于产品评论的情感分析和基于新闻评论的情感分析；按其研究的任务类型，可分为情感分类，情感检索和情感抽取等子问题[2]。

文本情感分析的基本流程如图 1 所示，包括从原始文本爬取，文本预处理，语料库和情感词库构建以及情感分析结果等全流程。由于文本原始素材爬取，分词等预处理技术已比较成熟，本文接下来将通过情感分析的主要任务情感分类，情感检索，情感抽取问题来分析和阐述已有的相关研究工作。

2 情感分类

情感分类又称情感倾向性分析，是指对给定的文本，识别其中主观性文本的倾向是肯定还是否定的，或者说是正面还是负面的，是情感分析领域研究最多的。

通常网络文本存在大量的主观性文本和客观性文本。客观性文本是对事物的客观性描述，不带有感情色彩和情感倾向，主观性文本则是作者对各种事物的看法或想法，带有作者的喜好厌恶等情感倾向。情感分类的对象是带有情感倾向的主观性文本，因此情感分类首先要进行文本的主客观分类。文本的主客观分类主要以情感词识别为主，利用不同的文本特征表示

收稿日期：2014-04-04
通信作者：宫玉龙(1985—)，男，硕士研究生。研究方向：文本情感分析。E-mail：mali@xupt.edu.cn

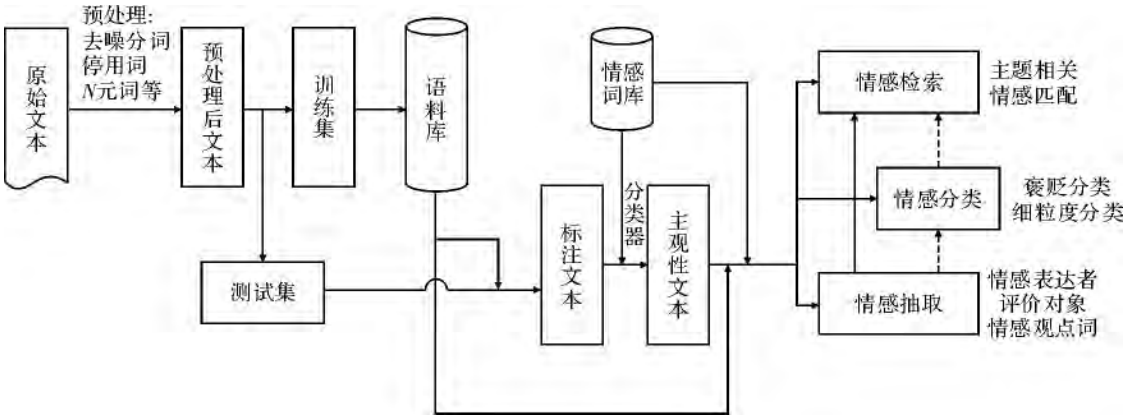

图1 文本情感分析基本流程

方法和分类器进行识别分类,对网络文本事先进行主客观分类,能够提高情感分类的速度和准确度[3]。

纵观目前主观性文本情感倾向性分析的研究工作,主要研究思路分为基于语义的情感词典方法和基于机器学习的方法。

2.1 基于语义的情感词典方法

(1)构建词典。情感词典的构建是情感分类的前提和基础,目前在实际使用中,可将其归为4类:通用情感词、程度副词、否定词、领域词。目前国内外,情感词典的构建方法主要是利用已有电子词典扩展生成情感词典。英文方面主要是基于对英文词典WordNet的扩充,Hu和Liu[4]在已手工建立种子形容词词汇表的基础上,利用WorldNet中词间的同义和近义关系判断情感词的情感倾向,并以此来判断观点的情感极性。中文方面则主要是对知网Hownet的扩充,朱嫣岚[5]利用语义相似度计算方法计算词语与基准情感词集的语义相似度,以此推断该词语的情感倾向。此外,还可以建立专门的领域词典,以提高情感分类的准确性。

(2)构建倾向性计算算法。基于语义的情感词典的倾向性计算不同于所需大量训练数据集的机器学习算法,主要是利用情感词典及句式词库分析文本语句的特殊结构及情感倾向词,采用权值算法代替传统人工判别或仅利用简单统计的方法进行情感分类。给情感强度不同的情感词赋予不同权值,然后进行加权求和。文献[6]利用加权平均算法式(1)计算,可有效提高通用领域情感分类的效率和准确率

$$\overline{E} = \frac{\sum_{i=1}^{N_p} \mathrm{wp}_i + \sum_{j=1}^{N_n} \mathrm{wp}_j}{N_p + N_n} \quad (1)$$

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。

(3)确定阈值来判断文本倾向性。一般情况下,加权计算结果为正是正面倾向,结果为负是负面倾向,得分为零无倾向。所得结果评价一般采用自然语言中经常使用的正确率、召回率和F值来评判算法效果。

基于情感词典的方法和基于机器学习的分类算法相比,虽属于粗粒度的倾向性分类方法,但由于不依赖标注好的训练集,实现相对简单,对于普遍通用领域的网络文本可有效快速地进行情感分类。

2.2 基于机器学习的情感分类方法

文本情感倾向性分析与传统的基于主题的文本分类相似但有所不同,基于主题的文本分类是把文本分类到各个预定义的主题上,如军事,互联网,政治,体育等,而情感分类不是基于内容本身的,而是按照文本持有的情感、态度进行判断。现有任何机器学习的分类方法都可以用到情感分类中来。基于机器学习的情感分类,其大致流程如下:首先人工标注文本倾向性作为训练集,提取文本情感特征,通过机器学习的方法构造情感分类器,待分类的文本通过分类器进行倾向性分类。

常用的情感分类特征包括情感词,词性,句法结构,否定表达模板,连接,语义话题等[7],研究者通过挖掘各种不同的特征以期望提高情感分类的性能。常用的特征提取方法有信息增益(Information Gain,IG)、CHI统计量(Chi-square,CHI)和文档频率(Document Frequency,DF)等。常用的分类方法有中心向量分类方法、K-近邻(K-Nearest-Neighbor,KNN)分类方法、贝叶斯分类器、支持向量机、条件随机场、最大熵分类器等。

最早从事情感分析研究的Pang等人[8]使用词袋(Bag-of-Feature)框架选定文本的N元语法(N-Gram)和词性(POS)等作为情感uo特征,使用有监督的机器学习的方法将电影评论分为正向和负向两类,分别使用朴素贝叶斯,最大熵模型和支持向量机作为有监督学习算法的分类器。结果显示支持向量机在几种分类方法中效果最好,分类准确率达到80%。文本

情感分类的准确率难以达到普通文本分类的水平,主要是情感文本中复杂的情感表达和大量的情感歧义造成的。

在基于机器学习的情感分类算法中,每篇文章被转换成一个对应的特征向量来表示。特征选择的好坏将直接影响情感分析任务的性能。在 Pang 等人的研究基础上,后续研究主要是把情感分类作为一个特征优化任务[9-11]。随着语义特征信息的加入和训练语料库的发展,基于机器学习的分类将会有广阔的发展前景。

3 情感检索

情感检索是从海量文本中查询到观点信息,根据主题相关度和观点倾向性对结果排序。情感检索返回的结果要同时满足主题相关和带有情感倾向或指定的情感倾向,是比情感分类更为复杂的任务。主题相关度和观点倾向性对结果排序,

随着人们网络检索需求的增高,在传统搜索中加入情感倾向成了搜索技术中一个新的研究热点。和传统的互联网搜索相似,情感检索有两个主要任务:(1)检索和查询相关的文档或句子。(2)对检索的相关文档或句子进行排序。与传统搜索不同的是互联网搜索的任务只要求找到和查询相关的文档和句子,而情感检索还要确定文档和句子是否表达了观点,以及观点是正面的或是负面的。目前情感检索主要实现方法有两种:一是按传统信息检索模型进行主题相关的文档检索,对检索结果进行情感分类;另一种是同时计算主题相关值和情感倾向值进行检索。

第一种方法一般使用传统的检索模型以及较为成熟的查询扩展技术,然后用情感分类方法进行倾向性计算。文献[12~13]给出的情感检索系统是国际文本检索会议 TREC(Text Retrieval Evaluation Conference)博客观点搜索任务的优胜者,该系统分为两部分检索部分和观点分类部分。检索部分完成传统的信息检索任务,同时在处理用户查询时将用户查询中的概念进行识别和消歧义,对于每个搜索查询进行同义词扩展,使用概念和关键字针对扩展后的查询对每个文档计算一个相似度,查询的关键字和文档的相关度是这两种相似度的综合。观点分类部分使用监督学习的方法使用两个分类器将文档分为两个类别带观点和不带观点的,带观点的文档再分为正面,负面或者混合的观点。第一个分类器训练数据是从评价网站包括 rateilt-all. com 和 epinion. com 收集大量带观点的数据和从维基百科等客观性网站收集不带观点的训练数据。第二个分类器训练数据来自评论网站包含打分的评论,低的打分表明负面观点,高的打分表明正面观点。这里两种监督学习的分类器都采用支持向量机。在 TREC 博客检索数据集研究的基础上,研究者采用不同的情感分类方法开展了后续研究[14-16]。

上面的方法是将检索和情感分类独立计算的,实际中主题相关和情感匹配是有关联的,需要同时计算主题相关和情感匹配,这是因为不同的情感词在文档中对不同的查询词下可能有相反的情感倾向。第二种方法则是同时考虑主题相关和情感文档排序,选择排序策略时需要同时兼顾。很多学者[17-18]对排序策略进行了研究,一般是分别计算情感倾向值和查询相关度值,然后加权求和进行排序。Zhang 等人[19]提出一种融合文档情感得分和文档查询相关度得分的概率生成模型排序方法,取得了理想的效果。

情感信息检索是传统信息检索技术和情感分析技术的融合,如何更好的融合二者得到理想的情感检索结果是未来要重点关注的。

4 情感抽取

情感抽取是指抽取情感文本中有价值的情感信息,其要判断一个单词或词组在情感表达中扮演的角色,包括情感表达者识别,评价对象识别,情感观点词识别等任务。

情感表达者识别又称观点持有者抽取,其是观点、评论的隶属者。在社交媒体和产品评论中,观点持有者通常是文本的作者或者评论员,其的登录账号是可见的,观点持有者抽取比较简单。而对于新闻文章和其他一些表达观点的任务或者组织显式的出现在文档时,观点持有者一般则是由机构名或人名组成,所以可采用命名实体识别方法进行抽取。Kim[20]等人借助语义角色标注来完成观点持有者的抽取。然而这些处理方法会导致较低的语言覆盖现象和较差的领域适应性,可以通过基于模式识别的信息抽取(Information Extraction)和机器学习(Machine Learning)技术来解决[21]。

评价对象和情感词抽取在情感分析中具有重要作用。利用评价对象和情感词的抽取,可以构建领域相关的主题词表和情感词表,情感词表的构建在情感分类部分已做阐述。评价对象是指某段评论中的主题,是评论文本中评价词语修饰的对象,现有的研究大多将评价对象限定在名词或名词短语的范畴内,一般使用基于模板和规则的方法抽取评价对象。规则的制定通常基于一系列的语言分析和预处理过程,命名实体识别,词性标注和句法分析等方法[22-25]都被用来进行评价对象抽取。文献[26]便是使用 3 条限制等级逐

渐渐进的词性规则从评价对象集中抽取评价对象,取得了较好的结果。

情感抽取是情感分析的基础任务,通过对大量的情感文本分析,有价值的情感信息抽取对于情感分析的上层任务情感检索和情感分类有直接帮助,如何准确抽取情感信息一直都是研究者关注的重点。

5 文本情感分析评测

近年来,情感分析得到了越来越多研究机构和学者的关注,在 SIGIR、ACL、WWW、CIKM、WSDM 等著名国际会议上,针对这一问题的研究成果层出不穷[27],国内外研究机构组织了众多相关评测来推动情感分析技术的发展。

由国际文本检索会议 TREC 针对英文文本观点检索任务的博客检索任务(Blog Track),篇章情感分类任务,以及其他一些有趣的情感分析任务;由日本国立信息学研究所主办的搜索引擎评价国际会议 NTCIR(NII Test Collection for IR Systems)针对日、韩、英、中文文本的情感分类以及观点持有者抽取任务。由中文信息学会信息检索委员会主办的每年一次的中文倾向性分析评测 COAE(Chinese Opinion Analysis Evaluation)已举办了 5 届,在关注情感词语和观点句子的抽取以及倾向性识别的基础上重点对于否定句、比较句以及微博观点句进行评测[28]。

众多研究机构的评测推动了情感分析研究的发展,出现了很多有代表性的情感分析语料库资源,文献[29~30]对语料库构建进行了详细阐述,如康奈尔影评数据集(Cornell Movie - Review Datasets),多视角问答(Multiple - Perspective Question Answering,MPQA)语料库,TREC 测试集,NTCIR 多语言语料库(NTCIR multilingual corpus),中文 COAE 语料库等。

6 结束语

本文对文本情感分析国内外的研究进展进行了综述,重点对情感分析中的几个关键问题情感分类,情感检索,情感抽取进行了介绍。文本情感分析作为自然语言处理和文本挖掘的一个新的研究方向有很多值得深入研究的课题。从现阶段的研究分析看,未来需要深入研究的问题有以下几方面:(1)基础性问题,由于中文语言的复杂性,自然语言处理相关技术中如分词和词性标注,句法分析等都会影响文本情感分析的结果,应加强相关技术的研究,提高对文本和句子的语义理解,从而更好的进行文本情感分析工作;(2)基于微博等新社交媒体的情感分析,这些新媒体用户量大,时效性强,语言简短口语化,包括的领域更加广泛,数据量巨大,对其进行情感分析有新的挑战和意义;(3)面向应用的情感分析,现有的研究大多是粗粒度的情感分析,应该更精确地更细粒度地对某一个具体的评价对象进行分析,把情感分析与实际应用相结合,比如票房预测,股票预测等;(4)语料库和文本情感分析评测,情感词库和标注语料库是文本情感分析研究的基础,应研究构建标准的情感词库和语料库,完善评测标准,推动中文情感分析理论和技术的研究和应用。

参考文献

[1] PANG B,LEE L. Opinion mining and sentiment analysis [J]. Foundations and Trends in InformationRetrieval,2008,2(1 - 2):130 - 135.

[2] 赵妍妍,秦兵,刘挺,等. 文本倾向性分析 [J]. 软件学报,2010,21(8):1834 - 1848.

[3] 厉小军,戴霖,施寒潇,等. 文本倾向性分析综述 [J]. 浙江大学学报,2011,45(7):1167 - 1175.

[4] HU M,LIU B. Mining and summarizing customer reviews [C]. NY,USA:Proceedings of Knowledge Discovery and Data Mining,2004:168 - 177.

[5] 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.

[6] 张昊旻,石博莹,刘栩宏. 基于权值算法的中文情感分析系统研究与实现 [J]. 计算机应用研究,2012,29(12):4571 - 4573.

[7] 李方涛. 基于产品评论的情感分析研究 [D]. 北京:清华大学,2011.

[8] PANG B,LEE L,VAITHYANATHAN S. Thumbs up:sentiment classification using machine learning techniques [C]. PA,USA:Proceedings of the ACL - 02 Conference on Empirical methods in natural language processing - Volume 10,Stroudsburg,Association for Computational Linguistics,2002:79 - 86.

[9] MELVILLE P,GRYC W,LAWRENCE. Sentiment analysis of blogs by combining lexical knowledge with text classification [C]. New York:Proceedings of SIGKDD,ACM,2009.

[10] LI S,HUANG C,ZHOU G. Employing personal impersonal views in supervised and semisupervised sentiment classification [C]. New York:Proceedings of ACL,ACM,2010:414 - 423.

[11] KUMAR A,SEBASTIAN T M. Sentiment analysis on twitter [J]. International Journal of Computer Science Issues,2012,9(4):628 - 633.

[12] ZHANG W,YU C,MENG W. Opinion retrieval from blogs [C]. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management,ACM,2007:831 - 840.

[13] ZHANG W,JIA L,YU C,et al. Improve the effectiveness of the opinion retrieval and opinion polarity classification [C]. MA USA:Proceedings of the 17th ACM Conference on Infor-

mation and Knowledge Management,ACM,2008:1415 – 1416.
[14] BALAHUR A,BOLDRINI E,MONTOYO A,et al. Opinion question answering:towards a unified approach [C]. ECAI,2010:511 – 516.
[15] MOGHADDAM S,ESTER M. AQA:aspect – based opinion question answering [C]. IEEE 11th International Conference on Data Mining Workshops,IEEE,2011:89 – 96.
[16] OH J H,TORISAWA K,HASHIMOTO C,et al. Why question answering using sentiment analysis and word classes [C]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics,2012:368 – 378.
[17] GARCÍA – MOYA L,ANAYA – SÁNCHEZ H,BERLANGA – LLAVORI R. Combining probabilistic language models for aspect – based sentiment retrieval [M]. Berlin Heidelberg:Advances in Information Retrieval,Springer,2012:561 – 564.
[18] DIETZ L,WANG Z,HUSTON S,et al. Retrieving opinions from discussion forums [C]. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management,ACM,2013:1225 – 1228.
[19] ZHANG M,YE X. A generation model to unify topic relevance and lexicon – based sentiment for opinion retrieval [C]. New York,NY,USA:Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM,2008:411 – 418.
[20] KIM S,HOVY E. Extracting opinions,opinion holders and topics expressed in onling news media text [C]. Text:Proceedings of the ACL Workshop on Sentiment and Subjectivity,2006:1 – 8.
[21] CARSTENS L,TONI F. Enhancing sentiment extraction from text by means of arguments [C]. Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining,ACM,2013:4.
[22] BAO L,HANKS S J,MCALLISTER I A,et al. Extraction and summarization of sentiment information:U. S. Patent 7,792,841 [P]. 2010 – 9 – 7.
[23] CHEN W,ZONG L,HUANG W,et al. An empirical study of massively parallel bayesian networks learning for sentiment extraction from unstructured text [M]. Berlin Heidelberg:Web Technologies and Applications,Springer,2011.
[24] O'NEIL J. Entity sentiment extraction using text ranking [C]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM,2012:1024 – 1024.
[25] YI J,NASUKAWA T,BUNESCU R. Sentiment analyzer:extracting sentiments about a given topic using natural language processing techniques [C]. Proceedings of the IEEE International Conference on Data Mining (ICDM),2013.
[26] ROUTRAY P,SWAIN C K,MISHRA S P. A survey on sentiment analysis [J]. International Journal of Computer Applications,2013(11):76 – 83.
[27] 中国科学院计算技术研究所,中国科学院自动化研究所,山西大学. 第五届中文倾向性分析评测(COAE13)大纲[EB/OL]. (2013 – 4 – 18)[2013 – 10 – 5] http://ccir2013.sxu.edu.cn/COAE.aspx.
[28] 杨立公,朱俭,汤世平. 文本情感分析综述[J]. 计算机应用,2013,33(6):1574 – 1578.
[29] LIU B,ZHANG L. A survey of opinion mining and sentiment analysis [M]. Springer US:Mining Text Data,2012:415 – 463.

(上接第 179 页)

[3] ZHU Q Y,YANG X F,YANG L X,et al. Optimal control of computer virus under a delayed model [J]. Applied Mathematics and Computation,2012,218(23):11613 – 11619.
[4] 邓兵,陶然,平殿发,等. 基于分数阶傅里叶变换补偿多普勒徙动的动目标检测算法[J]. 兵工学报,2009,30(10):1034 – 1039.
[5] 叶青,黄炎磊. 非均匀分布入侵检测模型的研究与仿真[J]. 科技通报,2013,29(8):169 – 171.
[6] 赵鹏军,邵泽军. 一种新的改进的混合蛙跳算法[J]. 计算机工程与应用,2012,48(8):48 – 50.
[7] 夏秦,王志文,卢柯. 入侵检测系统利用信息熵检测网络攻击的方法[J]. 西安交通大学学报,2013,47(2):14 – 19.
[8] 吴春琼. 基于特征选择的网络入侵检测模型[J]. 计算机仿真,2012,29(6):136 – 139.
[9] 王睿. 一种基于回溯的 Web 上应用层 DDOS 检测防范机制[J]. 计算机科学,2013,40(S2):175 – 177.