DOI: 10.3785/j.issn.1008-973X.2011.07.005

文本倾向性分析综述

厉小军[1], 戴 霖[1], 施寒潇[1], 黄 琦[2]
(1. 浙江工商大学 计算机与信息工程学院, 浙江 杭州 310018;
2. 浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

摘 要: 介绍文本倾向性分析的基本流程, 从主观性文本识别、文本倾向性分析方法、现有系统及评测方法、语料库建设 4 个方面对现有文本倾向性分析技术进行介绍和概括. 综述了文本倾向性分析的 3 类研究方法: 简单统计方法、机器学习方法和细粒度情感相关性分析方法, 分析这 3 类研究方法的特点, 从算法复杂性、效率和适用范围等方面比较各自的优缺点. 概括现有研究的成就和不足, 从基础性问题、具体应用的实现方法 2 个方面提出研究的前景.

关键词: 文本倾向性分析; 情感语料库; 主观性文本识别; 意见挖掘
中图分类号: TP 18; TP 391 文献标志码: A 文章编号: 1008-973X(2011)07-1167-08

Survey on sentiment orientation analysis of texts

LI Xiao-jun[1], DAI Lin[1], SHI Han-xiao[1], HUANG Qi[2]
(1. *School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou* 310018, *China*;
2. *College of Computer Science and Technology, Zhejiang University, Hangzhou* 310027, *China*)

Abstract: The basic flow of sentiment orientation analysis of texts was introduced, and the primary four aspects of current interesting researches were presented: subjectivity text recognition, sentiment orientation analysis method of texts, existing systems and evaluation methods, construction of corpus. Then three methods and their characteristics were summarized, i.e. simple statistics, machine learning and fine-grained sentiment relative analysis method. Merits and demerits of methods were analyzed from complexity, efficiency and applicable scope. Finally, the current achievements and shortages were summarized, and forecasted research perspectives were proposed including basic problem and implementation method of specific application.

Key words: sentiment orientation analysis of texts; sentiment corpus; subjective text recognition; opinion mining

随着 WEB2 0 技术的发展, 网络资源与日俱增, 越来越多的用户通过博客、评论网站、论坛等发表自己对一些事件、商品等的看法, 但仅仅通过人工浏览来获取大众观点信息是一件非常繁琐和困难的事情. 基于文本倾向性分析的意见挖掘技术应运而生. 基于文本倾向性分析的意见挖掘, 相对于主题挖掘, 需要对文本进行一定的智能化理解——倾向性分析, 在此基础上提取作者的意见、情感和态度等信息. 目前, 文本倾向性分析已成为自然语言处理领域的研究热点之一, 国内外越来越多的学者开始开展这方面的研究, 但还没有很完整的综述性文章. 姚天昉等[1]从观点分析的角度, 对文本意见挖掘进行综述; 周立柱等[2]从技术方面对情感分析研究进行综述. 本文在两者基础上, 考虑国内外最新进展, 对当前

收稿日期: 2010-05-18. 浙江大学学报(工学版)网址: www.journals.zju.edu.cn/eng
基金项目: 浙江省重大科技专项资助项目(2008C13082); 浙江省自然科学基金资助项目(Y1090688); 中央高校基本科研业务费专项资金资助项目(2009QNA5025).
作者简介: 厉小军(1974—), 男, 教授, 从事企业信息管理、自然语言处理研究. E-mail: lixj@zjgsu.edu.cn
通信联系人: 黄琦, 男, 副教授. E-mail: kylehq@163.com

文本倾向性分析研究现状进行概括和展望.本文首先介绍倾向分析的基本流程,然后分析主观性文本识别技术和倾向性分析技术的现有方法和研究趋势,最后介绍情感语料库的建设现状和相关评测技术.

1 文本倾向性分析基本流程

文本倾向性分析基本流程如图 1 所示,具体步骤如下:

1)原始素材的收集整理.一般采用爬虫工具定时进行材料搜集,例如:开源的 Java 爬虫软件有 heritrix 、nutch 等.

2)文本预处理.对收集来的素材进行噪音消除、标签过滤、分词等工作,为后续分析提供较好的原始分析文本.例如:具有良好容错性的网页分析工具 HtmlParser,分词软件有中国科学院计算机所研究室编写的 ICTCLAS.

3)主观性文本识别.利用事先建立好的语料数据库或分类器进行主客观文本识别,剔除一些不含情感元素的客观性文本,提高精确度.

4)文本倾向性判别.针对提取的主观性文本,结合语料库,采用简单统计方法或基于机器学习或基于相关性分析的方法判断主观性文本的褒贬倾向.

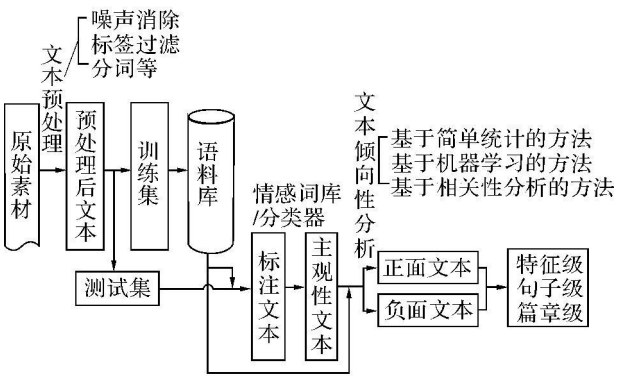

图 1 文本倾向性分析基本流程

Fig. 1 Basic flow of sentiment orientation analysis of texts

文本倾向性分析涉及自然语言处理、信息检索和抽取、机器学习、统计学、人工智能等多个领域,该流程涉及的学科较广泛.下述内容按照此流程进行展开,由于原始素材收集和文本预处理已比较成熟,接下来主要对主观性文本识别、文本倾向性分析和语料数据库建设进行阐述.

2 主观性文本识别

在通常的网络文本中,存在大量的客观性文本和主观性文本.客观性文本是一种不带有感情色彩的对个人、事物或事件的一种客观性描述,主观性文本主要描述作者对事物、人物、事件等的个人(或群体、组织等)想法或看法[1].主观性文本是文本倾向分析的主要对象,因此,对大量的网络文本事先进行主客观文本识别非常重要,能够有效地缩小分析范围,提高分析速度和精确度.

在主观句识别中,较简单的方式是通过对各种形容词的识别提取来判断句子的主客观性[3].Finn 等[4]研究主客观句分类,得出基于词性标注的特征选择方法比词袋效果好.对于人工标注语料,Wiebe 等[5]研究短语、句子和篇章层次,得出对于不同的标注者,存在标注判别的较大差异.Yu 等[6]对新闻这类主要讲"事实"的文本进行主客观句子识别,利用 SimFinder 工具计算句子相似度,构造训练集,结合各类词性信息构建贝叶斯分类器,提出多分类器的构建以解决训练集构造的不确定性和训练集质量的问题.对于主客观句子识别,比较常用的办法是结合词性标注,利用贝叶斯分类器进行分类[7-9].

Pang 等[10]利用属性相同的句子位置分布较近的特点,将候选句子构成一幅图,从而将主客观句分类转化为求图的最小割问题,实现 Cut-based 分类器,对主客观句进行分类识别.

在中文方面,叶强等[11]在 N-POS 语言模型的基础上,利用 CHI 统计方法提取中文词类组合模式,提出主观文本词类组合模式提取方法,建立中文双词主观情感词类组合模式——2-POS 模型,在语料试验中取得了良好的效果.

总之,主观性文本识别主要以情感词为主,利用各种文本特征表示方法和分类器(一般采用朴素贝叶斯分类器)进行分类识别.文本在经过主观性句子提取后,能够减少干扰,提高分类准确性.

3 文本倾向性分析方法

当前流行的文本倾向分析方法可以分为 2 个步骤[12].首先,对带有评价信息的词语、词组和语言模式等进行倾向判断,暂且将它们称为情感词、情感词组和情感语言模式,得出各个情感项的倾向权值;其次,利用不同的方法根据情感项的倾向值计算语义倾向评价,这种评价可以分为篇章级、句子级和特征级等.目前主要有 3 种研究思路,如表 1 所示.第 1 种是基于简单统计的倾向分类,对所有情感项进行简单的倾向性统计,统计方法有求和法和向量空间模型法,根据最终得分与事先设定阈值比较得出倾

表 1　不同文本倾向分析方法比较

Tab. 1　Comparison among different methods of sentiment orientation analysis

方法	优点	缺点	应用范围
基于简单统计的方法	实现简单	粒度粗	句子级、篇章级
基于机器学习的方法	知识获取客观，准确率较高	对训练集语料依赖性高	句子级、篇章级
基于相关性分析的方法	分析精准，粒度细	受限 NLP 技术或抽取算法	特征级

向评价，一般用于篇章级的倾向分类. 因为文本倾向分析基本都要涉及到统计方法，这里加了“简单”，以区别于其他方法. 第 2 种是基于机器学习的倾向分类，是采用基于机器学习的方式，通过对大量标注语料的训练生成倾向分类器，用来对测试文本进行分类. 目前主流的分类方法有支持向量机（support vector machine，SVM）、朴素贝叶斯（naïve Bayes，NB）和最大熵（maximum entropy，ME）等. 第 3 种是基于相关性的倾向分类，相对于前 2 种粗粒度的倾向分析，该方法属于细粒度分析. 它利用情感项和特征项的共现信息，或通过对情感项和特征项的句法依存分析，或通过语义角色标注，充分考虑情感项和特征项的相关性，实现基于特征级别的倾向分析.

3.1　情感项倾向值计算

情感项倾向值计算关键在于情感词的倾向计算. 在词语静态倾向值计算的基础上，考虑词语间动态组合产生的倾向变化，这涉及到情感词组和语言模式的倾向计算. 例如：“他们服务不够周全.”对“周全”进行情感词倾向计算为褒义，但由于前面的“不够”，使情感倾向发生了改变. 充分考虑词语间的相互作用，能够使倾向计算更有效.

词汇的倾向值计算，有基于 General Inquirer[13]、WordNet 和 HowNet 等知识库的方法. Kamps 等[14]利用 WordNet 中词语的同义结构图，通过待测词语与种子词的相似度计算得出待测词倾向值. 朱嫣岚等[15]在提取一定的基准词的基础上，利用 HowNet 提供的语义相似度和语义相关场功能计算待测词语与基准词的相似度来进行倾向值计算.

另外，文献[15]中基于语义相似度的词汇倾向值是待测词与各褒义基准词的相似度之和减去与各贬义基准词的相似度之和. 熊德兰等[16]利用各褒义基准词（或贬义基准词）构成多维正面空间（或反面空间），则待测词可用正面空间（或反面空间）的一点（或向量）表示，该点在基准坐标上的值即为与各基准词的相似度，待测词的倾向性可由该点与原点的长度（或向量长度）表示. 这种基于向量空间的词汇倾向计算法，比单纯的计算倾向和值更科学.

词汇倾向值计算，除了利用大型的知识库所提供的词语间关系外，还有利用词汇互信息（mutual information，MI）进行倾向计算. 代表者有 Turney 等[17-18]提出的利用 AltaVista 搜索引擎提供的 Near 操作，根据返回结果数得出待测词与基准词的共现概率，由此得出待测词与基准词的互信息. 待测词与各褒义基准词的点互信息（pointwise mutal information，PMI）之和减去与各贬义基准词的 PMI 之和，即为待测词语义倾向值.

由此可知，关于词语的倾向值计算，通常的方法是事先选取倾向较明显的词语作为基准词（一般以成对的形式出现，如（漂亮、丑陋））. 在此基础上，利用待测词与基准词的相关性进行倾向计算. 这种词间相关性可由知识库体现，也可由大量网络文本的统计分析体现.

另外，Hatzivassiloglou 等[19]考虑到形容词语义倾向受连接词“and”、“but”等约束，可以通过其中一个词的语义倾向推测出另一个. 如：由“beautiful and Y”可以推测出 Y 是褒义的. 考虑到连接词所揭示出的语义倾向关系，通过一定的聚类算法可以得出褒义形容词集和贬义形容词集. Esuli 等[20]通过对从词典中获得的词语的注释进行训练和分类来判断其他词语的语义倾向. Das 等[21]提出一种对带有否定词的情感项进行处理的方法，即对否定词影响域内词语加“_N”后缀，但由于涉及的无关特征项过多，导致了精确度的不理想.

情感词组[22]和语言模式[23]的倾向计算体现了情感词汇与修饰语间动态关系对倾向值的影响，计算难度更大的同时更实用.

3.2　基于简单统计的倾向分类

基于简单统计的倾向分类，目前的方法主要是通过适当的词语提取和倾向计算，对倾向值进行简单统计求得文本的整体倾向度[17,24]. Tsou 等[25]通过计算词语的语义倾向，综合考虑极性元素分布、密度和语义强度对新闻文本语义倾向进行统计，衡量公众对名人的评价.

基于简单统计的倾向分类，不仅可以采用简单统计求和方法实现，而且可以通过建立类别空间模型实现[26]. 基于简单统计的倾向分类虽然属于粗粒度的倾向分类，但由于实现简单、有一定的准确度，在倾向研究初期占据了一定的分量.

3.3　基于机器学习的文本倾向分类

基于机器学习的文本倾向分类流程大致如下：先对文本倾向性进行人工标注，提取文本特征表示，

并将其作为训练集,通过机器学习的方法构造分类器.待测文本可以通过分类器得到文本倾向性类别信息.常用的特征表示方法有:n-gram 特征表示、评价词组特征表示和单个词语特征表示等.常用的特征提取方法有:MI、信息增益(IG)、CHI 统计量(CHI)和文档频率(DF)等.常用的分类方法有:中心向量分类法、KNN 分类法、感知器分类法、贝叶斯分类法、最大熵分类法和支持向量机分类法等[27].Pang 等[28-32]在这方面都有一定的研究.

Pang 等[28-29]采用标准词袋技术和朴素贝叶斯、最大熵、SVM 分类方法,对 Usenet 上的电影评论进行文本倾向分类,并将它们和手工分类结果进行比较.实验结果显示,SVM 在几种分类方法中效果最好,分类准确率最高达到约 80%.Whitelaw 等[22]关注研究带评价信息的词组(appraisal groups),主要研究其中的形容词词组和修饰语(如"very good"或"not terribly funny"),提取带形容词的词组作为特征,结合标准词袋特征表示,基于这些特征,用向量空间模型表示文本,采用 SVM 对电影评论进行分类,达到了 90.2%的准确度.

在国内,马海兵等[33]结合传统的文本分类方法和 KNN、SVM 等基于向量空间模型的方法,将情感词本身权重纳入文本特征维权值考虑范围.徐琳宏等[34]提出一种结合语义特征和机器学习的汉语文本极性自动识别机制.首先通过 HowNet 计算词汇倾向性,选择极性明显的词汇作为特征值,用 SVM 分类器分析文本的褒贬.为了提高分类准确度,考虑否定副词和程度副词对语义倾向的影响.徐军等[27]利用朴素贝叶斯和最大熵方法研究新闻及评论语料的情感分类,通过一系列的实验得出各种方法的优劣对比.唐慧丰等[27]针对分类技术中的关键技术如:特征表示、选择和文本分类方法进行对比实验,采用 BiGrams 特征表示方法、信息增益特征选择方法和 SVM 分类方法,在足够大的训练集和选择适当数量特征的情况下,情感分类能够取得较好效果.

基于机器学习的倾向分类方法关键在于特征信息的有效提取,随着语义特征信息的加入和训练语料库的发展,基于机器学习的倾向分类将会有广阔的发展前景.

3.4 基于相关性的文本倾向分类

基于相关性的文本倾向分类为文本倾向分类提供了一种更细粒度的方式,它比基于篇章和句子的倾向分析更细化和精确.基于相关性的文本倾向分类由于充分考虑了情感词或词组与特征词的依存关系,提供了一种新的细粒度的基于特征的倾向分类,而不仅仅只是情感词或语言模式情感倾向的加权统计.同时,基于相关性的文本倾向分类对自然语言处理技术提出了更高的要求,需要对情感词或情感词组与描述特征项进行依存分析.目前,研究者通过对句子的语义角色标注(SRL)以及对文本句法信息的结构化抽取来解决这个问题.

Popescu 等[36]的研究工作专注于上下文各成分之间的相互作用.在获取词汇极性的基础上,根据已设定的一些限制,对词语、主题、句子以迭代方式指派极性,直至极性不再改变为止.这种方法需要大量的人工信息干预以达到较高的准确率.

Hu 等[37-38]利用情感词与候选特征的共现关系,在 POS 标注的基础上,实现了从高频特征到相关情感词再到低频特征的依次提取.由于在抽取时考虑了特征词和情感词的依赖关系,能够很好地进行基于特征的倾向分析.Miao 等[39]对每个产品的评论进行基于四元组(title, help, date, R-content)的抽取,其中 title 是指评论的题目,help 是指认为该评论是有助于他们的顾客数量,date 是指评论发表的日期,R-content 是指顾客评论的一系列句子集,对于每个句子,采用 Liu[40]提出的方法,即将每个句子表示为一个三元组[特征(feature),倾向(sentiment),句子相关内容(S-content)].通过对评论信息的结构化抽取建立索引,结合时间信息、认为评论有用的人数和词频信息进行排名权重计算,得出褒贬评论的统计曲线,分别进行褒贬评论对比展示和基于时间序列的图形化展示.Miao 等[39]提出的方法实现了特征级别的倾向分析.

考虑到不同类型的句子表达情感的方式不同,一种倾向分析方法不可能对所有的句子类型都非常适用,Narayanan 等[41]提出一种专门针对条件语句进行倾向分析的思路方法.通过对条件语句表达方式类型进行分析研究,基于时态信息对条件句子进行类别标注,结合各类特征表示信息,提出基于各分句、基于结果句和基于整个句子的分类方式,通过实验验证了各个分类方式在不同特征表示项下的性能.实验证明,条件句中观点分析的主要部分在结果句中,基于结果句和基于整个句子的分类方式的性能优于基于各分句分类方式的性能.

刘永丹等[42]采用格语法作为语义分析的基础,对文本进行语法和语义分析,提取相应格,然后与事先建立好的基于语义的过滤模板(该模板对行为受体和行为主体进行考虑)进行匹配,通过匹配距离函数和匹配相关函数计算匹配模板相关度,累加匹配模板相关度,最后与阈值比较得到文本过滤结果.

姚天昉等[43]利用领域本体抽取语句主题以及它的属性，在句法分析的基础上，识别主题和情感描述项间的关系，最终决定语句中每个主题的极性. 采用基于经验的语言模式方法，提出一种改进后的 SBV 极性传递算法，考虑到主谓结构(SBV 结构)、动宾结构(VOB)和定中结构(ATT)极性对相关主题词的较准确有效的传递，结合情感词库的建立，实现了合理有效的倾向分析，但该算法没有考虑语气问题.

基于相关性的文本倾向分析，能够实现特征级别的倾向判断，与句子和篇章级别的倾向分析相比，更实用，难度也更大. 它对自然语言处理技术水平依赖较大，现有方法主要通过信息结构化抽取和语义分析标注等来实现.

4　语料数据库建设

当今计算机的智能化程度不能对词语、句子或者篇章进行全自动的原发的倾向判别，倾向性判断依赖于一定规模的标注语料库. 语料库的建设是对具有倾向性的词语、语言模式或者句子以特定的方式存储标注，为文本倾向性分析提供数据基础. 语料库建设包括情感词库建设、语言模板建设、领域语料库建设和领域特征库建设等.

情感词库是具有代表性的、情感倾向较明显的词语的集合. 从词库中的词出发探讨汉语的情感系统是研究汉语情感分类的一个可行的方法，可为情感研究的分类提供很好的参照，也为文本情感信息提取的研究奠定了基础[44]. 情感词语并不局限于名词、形容词[45]，动词、状态词、声音词和成语也能表达丰富的情感体验；另外，副词中的否定副词和程度副词会对情感词语的倾向极性和程度产生一定影响. 一个全面而有效的情感词库不局限于某几种词类，应充分考虑语言的多样化.

在充分考虑情感词库涉及词类的基础上，情感词库的存储格式也至关重要，如情感词语与其倾向权值对应关系的建立以及如何设置科学有效的权值等. 目前，情感词库一般都伴随着某个分析方法的提出而建立，适用范围较狭隘，统一的标准的情感词库还处于待建状态. 现阶段有代表性的情感词库有“《知网》情感分析用词语集①”.

语言模板建设是指将一些特定的带有情感倾向的语言模式进行提取和标记，并将其作为倾向性分析的匹配模板[23]. 例如：“这朵花非常漂亮.”通过一定的句法分析和标注，提取出的语言模板为“(非常漂亮→)副词＋形容词”，对其进行一定的权值设定，得到此类模板对应的权值，当倾向分析时只需进行对应的模板匹配. 该方法需要大量的标注语言模板和机器学习，才能保证足够的精确度. 另外，孟凡博等[23]提出标注的格式设计问题，在考虑格式的统一标准和长度的基础上提出一种非常有效的五元标注格式.

相对于情感词库和语言模板的建设，领域语料库的建设特别是深度标注语料库建设困难得多. 领域语料库有原始语料素材库和标注语料库. 影评数据集作为使用较多的原始语料素材②，由电影评论组成，广泛应用于词汇和篇章情感倾向研究，但由于未进行细粒度的标注，应用范围有所限制. MPQA 库[46]是由 NRRC Summer Workshop 所开发、进行了深度标注的语料库，对论述持有者、对象、极性和强度等进行标注，缺点是规模过小. 语料库的标注程度和精确度直接影响倾向分析结果的准确度，语料库建设非常关键.

Quan 等[47]分别从篇章、段落和词汇 3 个级别对博客文本进行细粒度的标记，建立博客情感语料. 将情感分为 8 类，考虑情感主体与对象、情感词和短语、程度词否定词连词、各类修辞和标点符号(如：!)，该标注文档最后以 XML 形式输出. 实验证明，该语料能够有效地促进中文情感分析研究.

领域特征库建立涉及到本体建模准则及方法. 姚天昉等[48]建立的汽车本体，针对汽车的品牌和型号、汽车的机械部件、汽车的性能指标和汽车的总体评价 4 类特征定义了一个层次式的分类体系，以树的结构表示. 通过遍历树可以确定特征词之间的从属关系. 本体一般由领域专家建立，需要专家参与是目前本体的主要缺点之一，如何通过知识挖掘手段自动获取本体是目前和今后研究的重点[49]. 特征库的建立使文本倾向识别更加实用化.

5　现有系统和评测机构

随着研究的逐渐深入，对文本进行倾向分析的成型系统越来越多，如：Pulse[50]、Opinion Observer[51]、Sentiment Analyzer[52]、Web-Fountain[53]等. 其中，产品信息反馈系统 Opinion Observer 对网络顾客评价信息进行主观提取和特征提取，并对涉及产品特征的赞扬或批评评论进行统计，得出产品特

① http://www.keenage.com/html/c_bulletin_2007.htm

② http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/.

征综合质量的可视化结果. Sentiment Analyzer 系统通过建立情感词库和情感语言模式库,对在线评论进行特征术语提取、观点提取以及观点和特征关系的相关性分析,实现针对在线文本的基于特征的观点提取.在中文方面,姚天昉等[48]研究开发用于汉语汽车论坛的意见挖掘系统,可以实现在电子公告板、门户网站等各大论坛上的意见挖掘,对褒贬信息进行综合统计后给出可视化结果.

文本倾向性分析作为学术界的一个热点问题,近年来各类倾向分析或意见挖掘系统不断出现.为了使研究者对当前的研究现状有一个更直观的了解,从而更好地进行后续研究,评测工作不可小觑.

国际上针对文本倾向性分析的评测主要有文本检索会议(text retrieval conference, TREC)和多语言处理国际评测会议(NII test collection for IR system, NTCIR). TREC 主要面向英文 blog 语料的观点检索,要求在 blog 网页中检索出与给定主题相关并表达了观点的网页,同时要判别观点的倾向性. NTCIR 建立中、英、日 3 种语言的标准语料库,要求参加评测的系统能够进行句子主客观判别、句子倾向性判别、观点持有者判别、观点目标识别以及句子相关性判别.在国内,由中国中文信息学会信息检索专委会主办,2008 年推出了中文倾向性分析评测(chinese opinion analysis evaluation, COAE),对中文情感词的识别与分析、中文文本倾向性相关要素的抽取、篇章级中文文本主客观及倾向性判别进行评测.由评测结果可以得出,中文词语倾向性评测结果基本上能够达到满意的效果,评价对象抽取评测结果不理想,篇章级评测结果指标不高.

各类倾向性系统的开发成型以及权威评测机构的出现,在活跃学术氛围促进交流的同时,进一步推动了文本倾向性分析的研究进展和应用进程.

6 结 语

纵观文本倾向分析技术发展可知,自然语言处理技术、深度标注语料库建立以及全面、精准的分析方法是影响倾向分析技术发展的重要因素,总的来说,目前还存在不少困难.首先,由于网络文本的特殊性,需要对网络文本进行专门研究分析,得出表现类型、表达方式特别是网络新兴用语的研究报告.研究报告应以数据为本,理论分析为辅.其次,自然语言处理技术中的分词技术、词性标注、句法依存分析和语义标注等准确率直接影响倾向分析结果,所以这些方面亟需加强.再次,一个深度标注的统一标准的语料库的建立是非常有必要的,这将极大地促进广大研究者的研究进度.最后,高精确度的分析算法有待加强研究,希望通过各分析方法的取长补短实现突破.

从现阶段的研究情况分析来看,未来的研究趋势主要应针对两方面展开:

1)基础性问题.如情感词汇的获取及极性定量分析、特征识别.极性定量分析方面的研究由人工判定逐渐趋向于利用语料库进行词汇倾向相似度计算以及利用语言学知识,通过词的构成特点进行定量分析.从词汇的获取来看,未来的趋势是利用机器学习方法结合词汇所在的上下文环境来进行判断.特征识别未来的研究趋势将依赖于领域特征库的建设.

2)具体应用的实现方法.研究趋势是更多地引入自然语言处理技术,提高对句子、篇章的语义理解.同时利用语言学知识,针对特殊句型,研究语义多种表达以及语言修辞(如讽刺、反语)等,从而更好地挖掘文本中的倾向、观点等信息.

参考文献(References):

[1] 姚天昉,程希文,徐飞玉,等. 文本意见挖掘综述[J].中文信息学报,2008,22(3):71-80.
YAO Tian-fang, CHENG Xi-wen, XU Fei-yu, et al. A survey of opinion mining for texts [J]. **Journal of Chinese Information Processing**, 2008, 22(3): 71-80.

[2] 周立柱,贺宇凯,王建勇.情感分析研究综述[J].计算机应用,2008,28(11):2725-2728.
ZHOU Li-zhu, HE Yu-kai, WANG Jian-yong. Survey on research of sentiment analysis [J]. **Computer Application**, 2008, 28(11): 2725-2728.

[3] HATZIVASSILOGLOU V, WIEBE J M. Effects of adjective orientation and gradability on sentence subjectivity [C] // **Proceedings of the 18th Conference on Computational Linguistics**. USA: ACL, 2000: 299-305.

[4] FINN A, KUSHMERICK N, SMYTH B. Genre classification and domain transfer for information filtering [C] // **Proceedings of the 24th BCS-IRSG European Colloquium on Information Retrieval Research: Advances in Information Retrieval**. UK: Springer, 2002: 353-362.

[5] WIEBE J, BRUCE R, BELL M, et al. A corpus study of evaluative and speculative language [C] // **Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue**. USA: ACL, 2001: 1-10.

[6] YU H, HATZIVASSILOGLOU V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences [C] // **Proceedings of the 2003 Conference on EMNLP**. USA: ACL, 2003: 129-136.

[7] BRUCE R F, WIEBE J M. Recognizing subjectivity: a case study in manual tagging [J]. **Natural Language Engineering**, 1999, 5(2): 187-205.

[8] WIEBE J, RILOFF E. Creating subjective and objective sentence classifiers from unannotated texts [C] // **Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics**. Germany: Springer, 2005: 475-486.

[9] WILSON T, HOFFMANN P, SOMASUNDARAN S, et al. OpinionFinder: a system for subjectivity analysis [C] // **Proceedings of HLT/EMNLP 2005 Interactive Demonstrations**. Morristown, NJ, USA: ACL, 2005: 34-35.

[10] PANG B, LEE L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts [C] // **Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics**. Morristown, NJ, USA: ACL, 2004: 271-278.

[11] 叶强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007, 1(1): 79-91.
YE Qiang, ZHANG Zi-qiong, LUO Zhen-xiong. Automatically measuring subjectivity of chinese sentences for sentiment analysis to reviews on the internet [J]. **China Journal of Information Systems**, 2007, 1(1): 79-91.

[12] 来火尧, 刘功申. 基于主题相关性分析的文本倾向性研究[J]. 信息安全与通信保密, 2009(3): 77-78.
LAI Huo-yao, LIU Gong-shen. Prediction on semantic orientation of texts based on topic correlation [J]. **China Information Security**, 2009(3): 77-78.

[13] STONE P J, DUNPHY D C, SMITH M S, et al. **The general inquirer: a computer approach to content analysis** [M]. Cambridge, MA, USA: MIT, 1966: 1-6.

[14] KAMPS J, MARX M, MOKKEN R J, et al. Using WordNet to measure semantic orientations of adjectives [C] // **Proceedings of the 4th International Conference on Language Resources and Evaluation**. Lisbon, Portugal: [s.n.], 2004: 1115-1118.

[15] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006, 20(1): 14-20.
ZHU Yan-lan, MIN Jin, ZHOU Ya-qian, et al. Semantic orientation computing based on HowNet [J]. **Journal of Chinese Information Processing**, 2006, 20(1): 14-20.

[16] 熊德兰, 程菊名, 田胜利. 基于 HowNet 的句子褒贬倾向研究[J]. 计算机工程与应用, 2008, 44(22): 143-145.
XIONG De-lan, CHENG Ju-ming, TIAN Sheng-li. Sentence orientation research based on HowNet [J]. **Computer Engineering and Application**, 2008, 44(22): 143-145.

[17] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C] // **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Pennsylvania, USA: ACL, 2002: 417-424.

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

[19] HATZIVASSILOGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives [C] // **Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL**. Morristown, NJ, USA: ACL, 1997: 174-181.

[20] ESULI A, SEBASTIANI F. Determining the semantic orientation of terms through gloss classification [C] // **Proceedings of the 14th ACM International Conference on Information and Knowledge Management**. New York: ACM, 2005: 617-624.

[21] DAS S R, CHEN M Y. Yahoo! for amazon: extracting market sentiment from stock message boards[C] // **Proceedings of the 8th Asia Pacific Finance Association Annual Conference**. Bangkok, Thailand: [s.n.], 2001.

[22] WHITELAW C, GARG N, ARGAMON S. Using appraisal groups for sentiment analysis [C] // **Proceedings of the 14th ACM International Conference on Information and Knowledge Management**. New York: ACM, 2005: 625-631.

[23] 孟凡博, 蔡莲红, 陈斌, 等. 文本褒贬倾向判定系统的研究[J]. 小型微型计算机系统, 2009, 30(7): 1458-1462.
MENG Fan-bo, CAI Lian-hong, CHEN Bin, et al. Research on the recognition of text valence [J]. **Journal of Chinese Computer Systems**, 2009, 30(7): 1458-1462.

[24] NASUKAWA T, YI J. Sentiment analysis: capturing favorability using natural language processing [C] // **Proceedings of the 2nd International Conference on Knowledge Capture**. New York: ACM, 2003: 70-77.

[25] TSOU B K Y, YUEN R W M, KWONG O Y, et al. Polarity classification of celebrity coverage in the Chinese press [C] // **Proceeding of the 2005 International Conference on Intelligence Analysis**. Virginia, USA: [s.n.], 2005.

[26] 李艳玲, 戴冠中, 朱烨行. 基于类别空间模型的文本倾向性分类方法[J]. 计算机应用, 2007, 27(9): 2194-2196.
LI Yan-ling, DAI Guan-zhong, ZHU Ye-hang. Text tendency categorization method based on class space model [J]. **Computer Applications**, 2007, 27(9): 2194-2196.

[27] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88-94.

TANG Hui-feng, TAN Song-bo, CHENG Xue-qi. Research on sentiment classification of chinese reviews based on supervised machine learning techniques[J]. **Journal of Chinese Informatian Processing**, 2007, 21(6): 88-94.

[28] PANG B, LEE L, VAITHYANATHAN S. Thumbs up sentiment classification using machine learning techniques [C] //**Proceedings of the 2002 Conference on EMNLP**. Morristown, NJ, USA: ACL, 2002: 79-86.

[29] PANG B, LEE L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales[C] //**Proceedings of the ACL 2005**. Morristown, NJ, USA: ACL, 2005: 115-124.

[30] CHAOVALIT P, ZHOU L. Movie review mining: a comparison between supervised and unsupervised classification approaches [C] //**Proceedings of the 38th HICSS**. Big Island, Hawaii: [s.n.], 2005.

[31] MULLEN T, COLLIER N. Sentiment analysis using support vector machines with diverse information sources [C] //**Proceedings of the 2004 Conference on EMNLP**. Morristown, NJ, USA: ACL, 2004: 412-418.

[32] CHENG X. Automatic topic term detection and sentiment classification for opinion mining [D]. Master Thesis. Saarbrücken, Germany: The University of Saarland, 2007.

[33] 马海兵, 刘永丹, 王兰成, 等. 三种文档语义倾向性识别方法的分析与比较[J]. 现代图书情报技术, 2007, 23(4): 43-47.
MA Hai-bing, LIU Yong-dan, WANG Lan-cheng, et al. An analysis and comparison of three methods for document semantic orientation recognition[J]. **Modern Information Technology**, 2007, 23(4): 43-47.

[34] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100.
XU Lin-hong, LIN Hong-fei, YANG Zhi-hao. Text orientation identification based on semantic comprehension [J]. **Journal of Chinese Information Processing**, 2007, 21(1): 96-100.

[35] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95-100.
XU Jun, DING Yu-xin, WANG Xiao-long. Sentiment classification for chinese news using machine learning methods [J]. **Journal of Chinese Information Processing**, 2007, 21(6): 95-100.

[36] POPESCU A M, ETZIONI O. Extracting product features and opinions from reviews [C] //**Proceedings of HLT-EMNLP-05**. Morristown, NJ, USA: ACL, 2005: 339-346.

[37] HU Ming-qing, LIU Bing. Mining opinion features in customer reviews [C] //**Proceedings of 19th National Conference on Artificial Intellgience**. [S.l.]: AAAI, 2004: 755-760.

[38] LIU Bing, HU Ming-qing. Mining and summarizing customer reviews [C] //**Proceedings of the 2004 ACM SIGKDD International Conference**. New York: ACM, 2004: 168-177.

[39] MIAO Qing-liang, LI Qiu-dan, DAI Ru-wei. AMAZING: a sentiment mining and retrieval system [J]. **Expert Systems with Applications: An International Journal**, 2009, 36(3): 7192-7198.

[40] LIU Bing. **Web data mining**[M]. New York: Springer, 2006: 411-430.

[41] NARAYANAN R, LIU Bing, CHOUDHARY A. Sentiment analysis of conditional sentences [C] //**Proceedings of the 2009 Conference on EMNLP**. Morristown, NJ, USA: ACL, 2009: 180-189.

[42] 刘永丹, 曾海泉, 李荣陆, 等. 基于语义分析的倾向性文本过滤[J]. 通信学报, 2004, 25(7): 78-85.
LIU Yong-dan, ZENG Hai-quan, LI Rong-lu, et al. Polarity text filtering based on semantic analysis [J]. **Journal of China Institute of Communications**, 2004, 25(7): 78-85.

[43] 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报, 2007, 21(5): 73-79.
YAO Tian-fang, LOU De-cheng. Research on semantic orientation analysis for topics in chinese sentences [J]. **Journal of Chinese Information Processing**, 2007, 21(5): 73-79.

[44] 许小颖, 陶建华. 汉语情感系统中情感划分的研究[C] //第 1 届中国情感计算及智能交互学术会议. 北京: [s.n.], 2003: 199-205.
XU Xiao-ying, TAO Jian-hua. Research on the sentiment classification in chinese [C] //**Proceedings of the 1st Chinese Conference on Affective Computing and Intelligent Interaction**. Beijing: [s.n.], 2003: 199-205.

[45] 王治敏, 朱学锋, 俞士汶. 基于现代汉语语法信息词典的词语情感评价研究[J]. 中文计算语言学期刊, 2005, 10(4): 581-592.
WANG Zhi-min, ZHU Xue-feng, YU Shi-wen. Research on word emotional evaluation based on the grammatical knowledge-base of contemporary Chinese[J]. **International Journal of Computational Linguistics and Chinese Language Processing**, 2005, 10(4): 581-592.

[46] WIEBE J, BRECK E, BUCKLEY C, et al. NRRC summer workshop on multi-perspective question answering [R]. Bedford, MA: Northeast Regional Research Center, 2002.

[47] QUAN Chang-qin, REN Fu-ji. Construction of a blog emotion corpus for Chinese emotional expression analysis [C] //**Proceedings of the 2009 Conference on EMNLP**. Morristown, NJ, USA: ACL, 2009: 1446-1454.

(下转第 1186 页)

参考文献(References):

[1] 谭建荣, 顾新建, 祁国宁, 等. 制造企业知识工程理论、方法与工具[M]. 北京: 科学出版社, 2008: 318-341.

[2] 毕经元, 顾新建, 吕艳, 等. 基于知识元链接的汽车零部件知识管理系统[J]. 浙江大学学报: 工学版, 2009, 43(12): 2208-2212.
BI Jing-yuan, GU Xin-jian, LV Yan, et al. Auto-parts knowledge management system based on knowledge-unit linking [J]. **Journal of Zhejiang University: Engineering Science**, 2009, 43(12): 2208-2212.

[3] 顾新建, 祁国宁. 知识型制造企业[M]. 北京: 国防出版社, 2000: 193-214.

[4] 赵蓉英. 知识网络研究(II): 知识网络的概念、内涵和特征[J]. 情报学报, 2007, 26(3): 470-476.
ZHAO Rong-ying. Study on knowledge network (part II): the notion and characters of knowledge network [J]. **Journal of the China Society for Scientific and Technical Information**, 2007, 26(3): 470-476.

[5] 刘征, 孙守迁. 产品设计认知策略决定性因素及其在设计活动中的应用[J]. 中国机械工程, 2007, 18(23): 2813-2817.
LIU Zheng, SUN Shou-qian. Determined element of product design cognitive strategies and its applications in design [J]. **China Mechanical Engineering**, 2007, 18(23): 2813-2817.

[6] 胡恒杰, 顾新建, 暴志刚, 等. 面向行业网络的知识发现及共享服务平台研究[J]. 浙江大学学报: 工学版, 2008, 42(8): 1445-1451.
HU Heng-jie, GU Xin-jian, BAO Zhi-gang, et al. Research on industry oriented network knowledge discovery and sharing platform [J]. **Journal of Zhejiang University: Engineering Science**, 2008, 42(8): 1445-1451.

[7] TOLMAN E C. Cognitive maps in rats and men [J]. **Psychological Review**, 1948, 55(4): 189-208.

[8] JENKINS M, JOHNSON G. Entrepreneurial intentions and outcomes: a comparative causal mapping study [J]. **Journal of Management Studies**, 1997, 34(6): 895-921.

[9] NOH J B, LEE K C, KIM J K, et al. A case-based reasoning approach to cognitive map-driven tacit knowledge management [J]. **Expert Systems with Application**, 2000, 19(4): 249-259.

[10] KOSKO B. Fuzzy cognitive maps[J]. **International Journal of Machine Studies**, 1986, 24(1): 65-75.

[11] 倪旭东, 张钢. 作为思想挖掘工具的认知地图及其应用[J]. 科研管理, 2008, 29(4): 19-27.
NI Xu-dong, ZHANG Gang. Cognitive map used as an idea digging tool and its application [J]. **Science Research Management**, 2008, 29(4): 19-27.

(上接第 1174 页)

[48] 姚天昉, 聂青阳, 李建超, 等. 一个用于汉语汽车评论的意见挖掘系统[C] //中国中文信息学会 25 周年学术会议. 北京: 清华大学出版社, 2006: 260-281.
YAO Tian-fang, NIE Qing-yang, LI Jian-chao, et al. A opinion-mining system for chinese automobile review [C] //**Proceedings of the 25th workshop of Chinese Information Processing Society of China**. Beijing: Tsinghua University Press, 2006: 260-281.

[49] 邓志鸿, 唐世渭, 张铭, 等. Ontology 研究综述[J]. 北京大学学报: 自然科学版, 2002, 38(5): 730-737.
DENG Zhi-hong, TANG Shi-wei, ZHANG Ming, et al. Survey on Ontology research [J]. **Acta Scientiarum Naturalium Universitatis Pekinensis**, 2002, 38(5): 730-737.

[50] GAMON M, AUE A, CORSTON-OLIVER S, et al. Pulse: mining customer opinions from free text [C] // **Proceedings of the 6th International Symposium on Intelligent Data Analysis**. Madrid, Spain: Springer, 2005: 121-132.

[51] LIU Bing, HU Ming-qing, CHENG Jun-sheng. Opinion observer: analyzing and comparing opinions on the web [C] //**Proceedings of the 14th International Conference on World Wide Web**. New York: ACM, 2005: 342-351.

[52] YI J, NASUKAWA T, BUNESCU R, et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques [C] //**Proceedings of the 3rd IEEE International Conference on Data Mining**. Washington, DC: IEEE, 2003: 427-434.

[53] YI J, NIBLACK W. Sentiment mining in web-fountain [C] //**Proceedings of the 21th International Conference on Data Engineering**. Washington, DC: IEEE, 2005: 1073-1083.

[54] EGUCHI K, LAVRENKO V. Sentiment retrieval using generative models [C] //**Proceedings of the 2006 Conference on EMNLP**. Morristown, NJ, USA: ACL, 2006: 345-354.

[55] MISHNE G, GLANCE N. Predicting movie sales from blogger sentiment [C] //**Proceedings of the 21th National Conference on Artificial Intelligence**. Menlo Park, California: AAAI, 2006: 155-158.

[56] 王超, 李楠, 李欣丽, 等. 倾向性分析用于金融市场波动率的研究[J]. 中文信息学报, 2009, 23(1): 95-99.
WANG Chao, LI Nan, LI Xin-li, et al. The research on financial volatility with sentiment analysis [J]. **Journal of Chinese Information Processing**, 2009, 23(1): 95-99.