中文网络话题评论文本语义倾向分析

刘 姗，胡 勇
(四川大学信息安全研究所，四川 成都 610064)

[摘　要] 文本的情感倾向分析是一项具有较大实用价值的关键技术。文中主要针对短语级和句子级的情感倾向分析进行对比研究。采用情感粒度中的情感短语和情感句子模型，根据不同的短语搭配模式、语义依存关系方法的组合，对中文文本倾向性分析进行了研究。研究表明，采用选取合适短语搭配模式的方式，以情感句子为最小判断单位的倾向分析方法，并应用于网络话题的中文评论文本，能取得较好的倾向分析效果。

[关键词] 中文倾向性分析；语义技术；短语模式；情感句子判断
[中图分类号] TP393　　[文献标识码] A　　[文章编号] 1009-8054(2012)06-073-03

Analysis on Chinese Semantic Orientation Identification of Network Topic-oriented Comment Text

LIU Shan，HU Yong
(Institute of Information Security，Sichuan University，Chengdu Sichuan 610064，China)

[Abstract] Emotional-tendency analysis of the text is a key technology with fairly large practical value. This article mainly aims at the emotional -tendency analysis of the phrases and sentences level. By using the emotional phrase and sentence model and according to different phrases match modes and combination method of special semantic dependent relationship，the Chinese text tendentiousness is studied. Experiment indicates that the selection of appropriate phrases match mode and the tendentiousness analysis with emotional sentence as the smallest judgment unit，could achieve fairly good tendency analysis results when applied in the network topic-oriented comment text.

[Keywords] Chinese semantic orientation identification；semantic technology；phrase model；
emotional-sentence judgment

0　引言

随着网络的飞速发展，网上各种各样的文章和言论信息量越来越大，使文本倾向性分析逐渐成为了近几年热门的研究课题。目前，在这个领域中有几个主要的研究方向 [1-2]，其中观点提取和词汇倾向性分类为倾向分析核心技术，文本倾向性分类和主客观分类 [3] 也是倾向分析研究方向的重要分支。而关于情感分析中的研究思路主要为采用机器学习的方法、基于语义的方法和结合语义和机器学习的方法。文中侧重于对语义的方法进行倾向性研究。

文中主要针对短语级和句子级的情感倾向分析进行对比研究，采用情感粒度中的情感短语和情感句子模型，根据不同的短语搭配模式、语义依存关系方法的组合，对中文文本倾向性分析进行了研究。

收稿日期：2012-02-21
作者简介：刘姗，1988 年生，女，硕士，研究方向：信息安全；胡勇，1973 年生，男，副教授，硕士生导师，研究方向：信息安全。

1　倾向性分析

1.1　短语级别的倾向分析

1.1.1 短语的情感分析

在文本倾向分析中，仅仅依赖于单个情感词来判断文本倾向准确率不够高，会忽略一些重要的信息，比如情感词所针对的评价对象等重要因素。因此，仅仅从情感词上来分析所判断的文本倾向很不准确，为了解决这些问题，采用情感短语进行判断，利用情感词所在文本中的特征信息来进一步判断文本倾向 [4]。文中的组合短语的分析主要查找文本中能表达情感特征的短语。

1.1.2 短语模式的选取

利用短语作为文本特征来分析文本情感倾向，实现文本分类的研究。文献 [5] 中详细地做了不少模式分类的工作，在统计了 1 000 篇针对网络话题的评论性文本后，选取了如表 1 中出现频率较高的短语搭配模式。

1.1.3 短语模式倾向值计算

在文中所进行的针对网络话题的评论性文本的倾向判断中，短语模式在文本中出现的类型在一定时期内变化不大。在计算短语模式的情感倾向及其强度时，文中

根据短语模式分别出现在正向和负向情感倾向文本中的次数进行计算，即通过比较同短语模式在持有肯定情感倾向的文本中出现的次数和它在否定情感倾向的文本中出现的次数就可以确定该模式的情感倾向及其强度[6]。

由此，文中抽取了 1 000 篇热点新闻话题的帖子，正向和负向文本各 500 篇，得到短语模式的出现次数后进行了相应的计算，这样就得到了文本中每个短语的情感倾向值。最后对所有短语的情感倾向值进行加和，作为所判断文本的最终情感倾向值。

表 1 文中选取的短语搭配模式

短语搭配模式	在所判断文本中出现的频率/(%)
评价对象 + 形容词／名词	76
评价对象 + 否定词 + 形容词／名词	57
评价对象 + 副词 + 形容词／动词	54
评价对象 + 否定词 + 副词 + 形容词／名词	43
评价对象 + 副词 + 动词	23
副词 + 动词 + 评价对象	35
否定词 + 副词 + 动词 + 评价对象	21

1.2 句子情感倾向性分析

1.2.1 句子情感倾向性分析相关

包含情感的句子是段落级或者篇章级情感倾向分析的重点，着重分析这些句子有助于分析整个文本的情感倾向[7-9]。句子级的情感倾向分析主要是对句中的各种主观性信息进行分析。文中句子级的情感分析采用了语义分析的方法，对中文长短句和句子中词之间的依存关系进行分析，从而计算文本的情感倾向。语义分析包括句式分析和词的依存关系分析。

对中文句式的结构进行分析，中文句式的分类有：

1) 按结构分为单句、复句。

2) 按语气分为陈述句 (肯定句、否定句、双重否定句)、祈使句、感叹句、疑问句。

文中对句子的情感倾向分析采用对词的情感分析和对句式的情感分析来求解。

1.2.2 句式分析

对于句子的情感计算，首先分析句式，对于不同句式采用不同的处理方法，如表 2 所示 (P 为所判断句子的情感倾向)。

表 2 句式分析

句式	句子情感倾向	句式	句子情感倾向
并列	$P=P$	假设	无
递进	$P=P+0.5$	条件	$P=P$
转折	$P=-P$	因果	$P=P$
选择	$P=P$	否定	$P=-P$

1.2.3 依存关系分析

在以句子为单位的情感倾向判别过程中，对于有情感倾向的词，文中利用语义依存关系[6]分析模型来进行相应的判断。所判断句子的情感倾向可以定义为各个部分依存结构的合并操作概率乘积。最后，句式分析和依存关系分析的结果的乘积即为文本中有情感倾向的句子的情感倾向值，文本的情感倾向值即为有情感倾向的句子的情感倾向值的加和值。

1.3 短语和句子模式结合的倾向判断

文中所述的短语级和基于句子级的文本倾向处理方法的正确率和召回率还有待提高，基于 100 篇样本文本倾向分析：短语级的文本倾向处理正确率只有 86.5%，召回率只有 85.3%；而句子级的文本倾向处理正确率只有 84.7%，召回率 84.6%。因此，文中提出了利用短语级处理模式和句子级处理模式结合的判断模式进行句子倾向性分析。短语模式和句子模式结合的倾向判断模式的流程如图 1 所示。

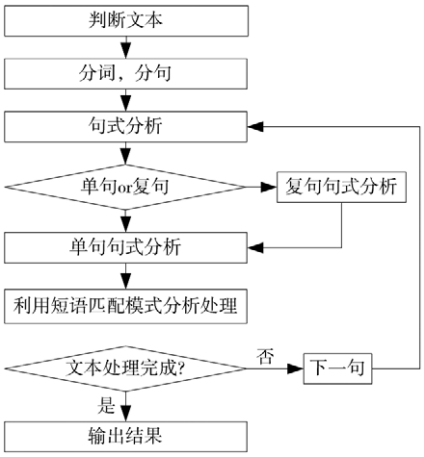

图 1 判断流程

这种倾向判断模式将上述的短语模式和句子判断模式相结合，具体描述如下：

1) 对网络话题的评论进行分词、分句，获得结果序列。

2) 对分句结果序列进行扫描，进行句式分析，并做标记。

3) 针对每句话，对分词后的结果提取短语匹配模式，利用合适的短语搭配模式进行情感短语的倾向判断，并对句子中情感短语的情感值累加后和句式判断的情感值进行相乘，最终获得句子的倾向值，根据句子倾向值最终确定文本的倾向值。如式 (1) 所示：

$$P(i)=S(i)\times\sum_{k=1}^{m_k}d_k \tag{1}$$

其中，S 为情感句子的句式分析结果；d_k 为情感句子中的情感短语的情感值；m_k 为情感句子中含有的情感短语的数量。

2 实验结果及分析

文中所用的情感词词表中情感词的主要来源是知网 (How Net) 发布的“情感分析用词语集”[10-11]，实验所用语料来自于大型网站。该语料选取了 600 篇文本，其中正向文本为 381 篇，标记为 Positive，占总文本数的

63.5%；负向文本 219 篇，标记为 Negative，占总文本数的 36.5%；语料长度字数 1 至 200 字不等。

实验结果中采用查准率 S(正确识别的篇数除以测试语料总篇数)，召回率 R(正确识别的篇数除以测试语料中总共正确的篇数) 来评测分析情感倾向的性能。3 种处理方法的实验结果如表 3 所示。

表 3 实验结果

语料类型	短语级 S/(%)	短语级 R/(%)	句子级 S/(%)	句子级 R/(%)	短语模式和句子模式结合 S/(%)	短语模式和句子模式结合 R/(%)
正向文本	87.3	84.56	85.11	82.33	95.53	96.21
负向文本	86.23	84.78	81.23	80.14	90.09	89.97

由实验结果可知：

1) 相对而言，句子级情感倾向分析法用于文本情感分类效果不如短语级情感倾向分析法。主要原因在于文本的情感类别更多地涉及人的主观心理因素，作者在用自然语言表达其情感时微妙复杂。

2) 现有的文本倾向分析中的句子倾向分析中，准确率和召回率在 88%左右，所以短语模式和句子模式结合的倾向判断模式有效地提高了本实验效果。

3) 召回率方面，负向文本的召回率明显没有正向文本的高，原因是在针对网络话题的评论文本中贬义词所在的上下文中含有更多的修饰副词，而且有些词的动态属性 [12] 使分析比较复杂。另一方面，也说明了中文表达的多样性，单纯用算法来识别也判断不出正确的极性，影响了负极性的召回率。

3 结语

文中主要针对短语级和句子级的情感倾向分析进行对比研究，采用情感粒度中的情感短语和情感句子模型，根据不同的短语搭配模式、语义依存关系方法的组合，对中文文本倾向性分析进行了研究。

结论表明，采用选取合适短语搭配模式的方式，以情感句子为最小判断单位的倾向分析方法，并应用于网络话题的中文评论文本，能取得较好的倾向分析效果。倾向性识别技术的研究目前处于起步阶段，还有很多问题有待研究，文中主要考察了句子和短语级的倾向分析方法，但还不够全面，在以后的实验中，可以将段落级倾向分析加入到针对网络话题的评论性文章的判断中，进一步加以研究。

参考文献

[1] TANG Huifeng，TAN Songbo，CHEN Xueqi. A Survey on Sentiment Detection of Reviews[J]. Expert Sys-tems With Applications，2009，36：10760-10773.

[2] 陈铭，李生红，陈秀真 . 基于句式结构的评论倾向性识别方法 [J]. 通信技术，2011，44(2)：100-101.

[3] 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(11)：79-91.

[4] 李钝，曹付元，曹元大，等 . 基于短语模式的文本情感分类研究 [J]. 计算机科学，2008，35(41)：132-134.

[5] 宋光鹏 . 文本的情感倾向分析研究 [D]. 北京：北京邮电大学，2008.

[6] 郭叶 . 中文句子情感倾向分析 [D]. 北京：北京邮电大学，2010.

[7] 李纲，程洋洋，寇广增 . 句子情感分析及其关键问题 [J]. 图书情报工作，2010，54(11)：104-127.

[8] 朱杰，刘功申，陈卓 . 中文文本倾向性分类技术比较研究 [J]. 信息安全与通信保密，2010(4)：56-58.

[9] 李海燕，李生红，张月国 . 面向离散文本舆情分析的分聚类方案 [J]. 信息安全与通信保密，2010(2)：65-67.

[10] 中文知网 . 情感分析用词语集 [EB/OL]. [2010-05-10]. http：//www.keenage.com/html/c_index.html.

[11] 蔺璜，郭姝慧 . 程度副词的特点范围与分类 [J]. 山西大学学报，哲学社会科学版，2003，26(2)：71—74.

[12] 姚天昉，娄德成 . 汉语情感词语义倾向判别的研究 [C]//2007 中文信息处理国际会议 (ICCC2007) 论文集 . 武汉：[出版社不祥]，2007：221-225.

本期简讯 Brief News In This Lssue

《信息安全与通信保密》杂志启用科技期刊学术不端文献检测系统

为了提高来稿质量，杜绝学术造假，促进《信息安全与通信保密》的健康发展，从 2009 年 1 月起，本刊编辑将正式启用科技期刊学术不端文献检测系统，对所有来稿进行检查。对于检测出有不端行为的稿件，编辑部将直接退稿。在此，希望广大作者在撰写论文时，一定要本着实事求是的科学精神，引用他人的研究成果时务必在参考文献中列出，并在正文中相应位置进行标注。大家共同努力，维护学术研究的诚信，杜绝学术不端行为，促进《信息安全与通信保密》的可持续发展，为广大作者搭建一个更好、更高、更权威的学术争鸣和技术交流的平台。

《信息安全与通信保密》杂志社

二 OO 九年一月一日