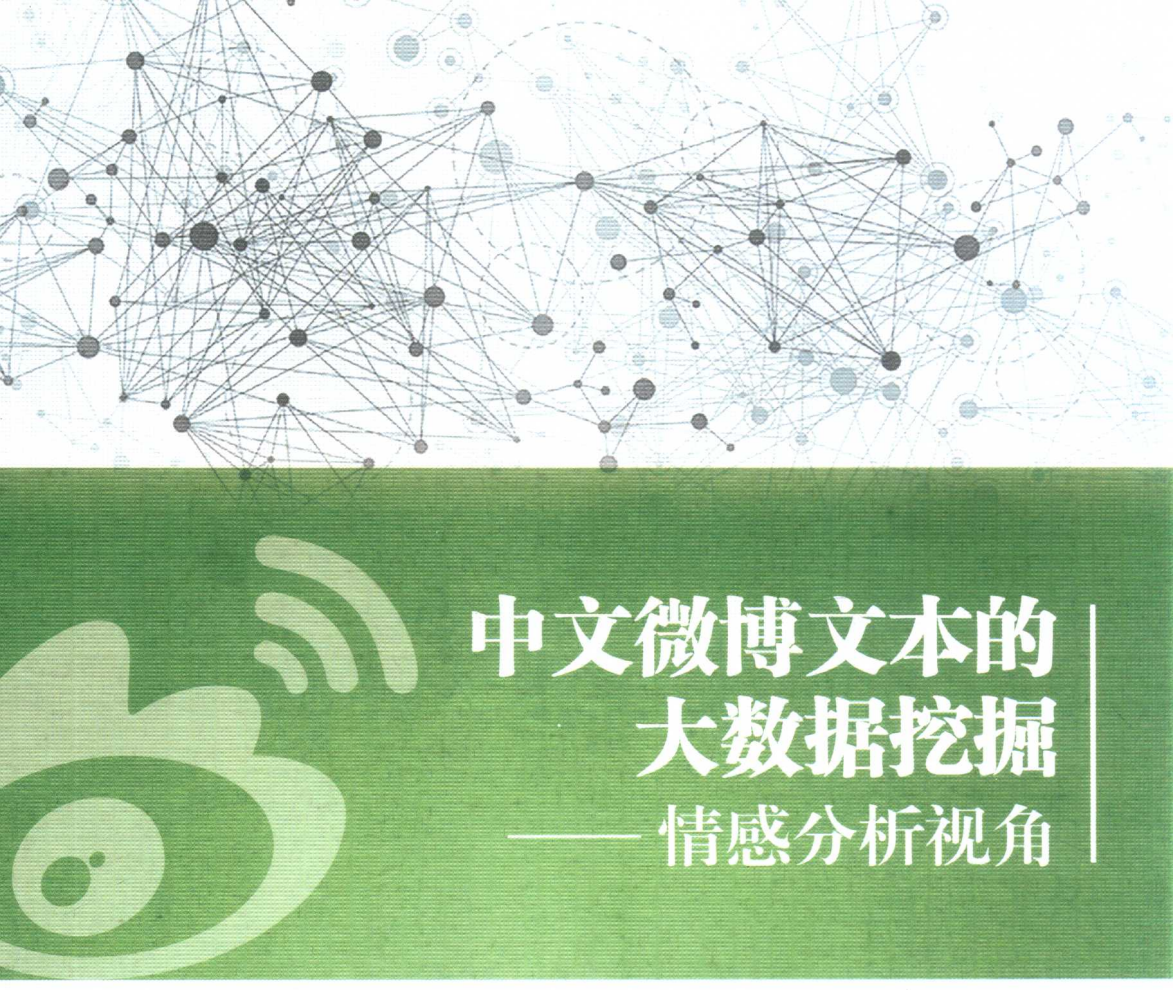
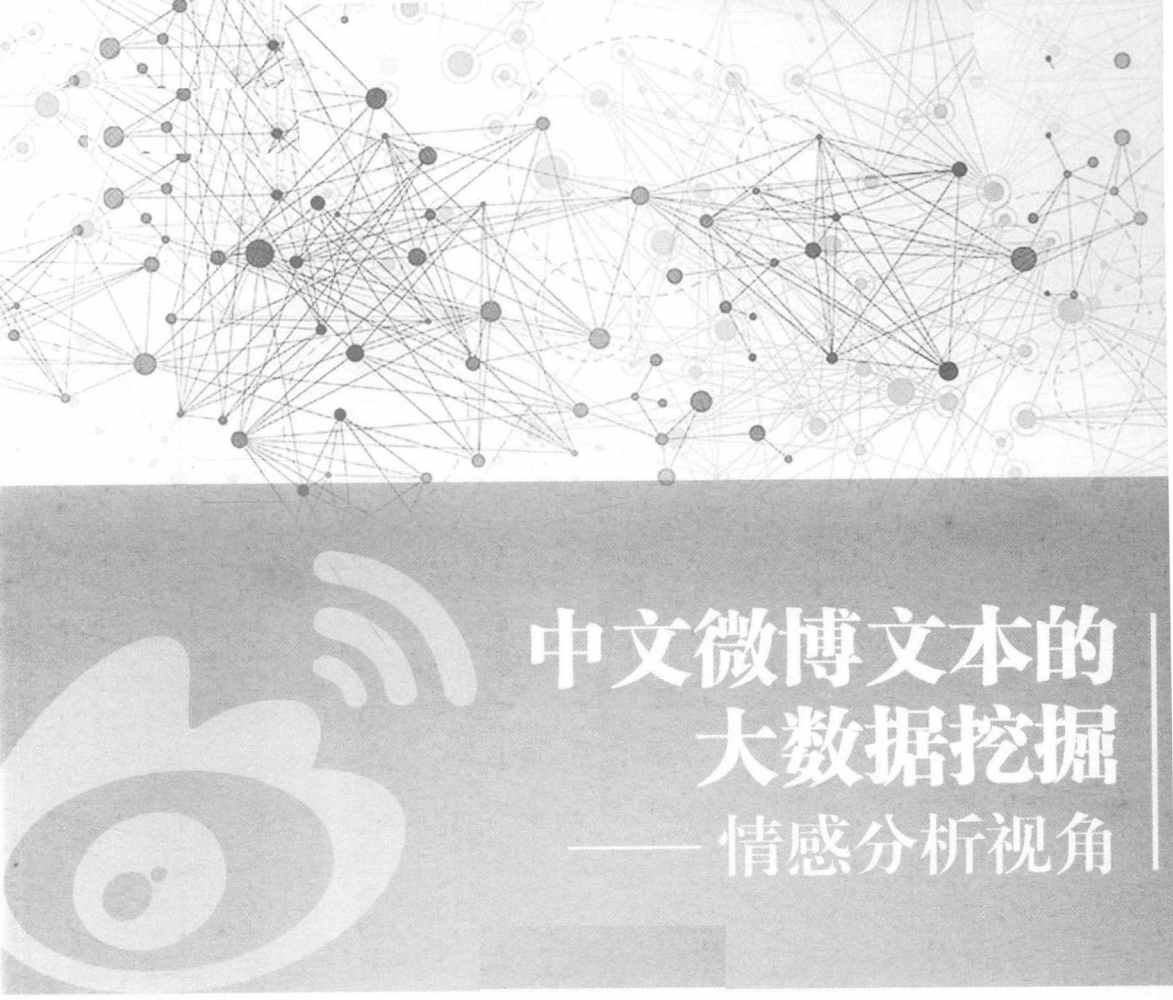


中文微博文本的大数据挖掘
——情感分析视角

史　伟◎著

中国社会科学出版社



中文微博文本的大数据挖掘
——情感分析视角

史　伟◎著

中国社会科学出版社

图书在版编目（CIP）数据

中文微博文本的大数据挖掘：情感分析视角/史伟著．—北京：中国社会科学出版社，2017.11

ISBN 978-7-5161-9312-9

Ⅰ.①中… Ⅱ.①史… Ⅲ.①互联网络—传播媒介—数据处理 Ⅳ.①G206.2

中国版本图书馆 CIP 数据核字(2016)第 270757 号

出 版 人 赵剑英
责任编辑 卢小生
责任校对 周晓东
责任印制 王 超

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http：//www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京明恒达印务有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2017 年 11 月第 1 版
印 次 2017 年 11 月第 1 次印刷

开 本 710×1000 1/16
印 张 13
插 页 2
字 数 181 千字
定 价 56.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683
版权所有 侵权必究

前　言

近年来，随着社交网络、电子商务和移动互联网的迅猛发展，人类社会数据的快速增长给许多行业带来了共同面对的严峻挑战和宝贵机遇，因而信息社会已经进入了大数据（BigData）时代。其中，互联网大数据的涌现不仅改变了人们的生活与工作方式、企业的运作模式，甚至还引起科学研究模式的根本性改变。尤其是随着Web2.0时代的到来，越来越多的人愿意在线表达自己的心情（微博）、发表自己对于政策的看法（新闻评论）、发布自己对于产品的评价（产品评论），等等。区别于传统结构化的数据，互联网大数据的表现形式大多为非结构化或半结构化的评论文本形式，对这些数据的挖掘和分析工作显得更加棘手。情感分析技术的出现正好满足了人们对大规模数据进行观点分析的需要。

情感分析，又称倾向性分析和意见挖掘，是情感计算的重要分支，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。目前，情感分析技术，主要包括机器学习方法及语义方法两类。基于机器学习的情感分类方法需要用大量的训练样本对分类模型进行训练，且训练样本集的建立需要采用人工方法进行手工标志。因此，近年来研究者将情感分析研究集中在对训练样本的需求量较低的语义方法上。

笔者长期从事基于语义的情感分析、文本挖掘、大数据研究工作。博士学习期间的主要研究围绕本体结构的描述→情感本体的构建→在线评论的语义分析→在线评论的情感分析→情感分析的应用，逐步深入展开。研究的载体主要基于中文微博文本展开，之所

以选择微博文本作为研究对象，是因为微博越来越成为大数据时代新媒体的核心载体，人们在微博平台发表对各种产品或服务的感受，表达对各种事件的观点，微博中蕴含着极其丰富的用户情感信息，而且微博文本属于典型的评论文本，数据资源对外公开，获取也较为方便，对微博文本进行情感分析具有重要的应用价值和实际意义。

基于先前的研究基础，本书从情感分析视角对中文微博文本的大数据挖掘进行研究，是对博士期间工作的深入。对微博中大数据，大致可以分为三类：一是结构化数据，用户注册微博时所提交的个人信息、关注和被关注等所产生的数据；二是半结构化数据，如用户发布的信息、用户之间互动（转发、评论、私信）的信息等所产生的数据；三是非结构化数据，如用户的情感观点和演化、在微博中留下的现实踪迹和发展、用户之间互动等信息数据。对不同的数据类型，需要采用不同的研究方法，对微博中的大数据进行挖掘。本书的研究脉络也基本按照对微博中不同类型数据的分析来展开。

第一章主要介绍互联网大数据、情感分析及微博大数据研究等问题的背景，以及目前在情感分析技术、微博挖掘和本体建模理论等方面的研究现状，进一步明确了本书研究的目的和价值。

第二章对本书研究涉及的相关领域的工作进行了综述。主要从基于语义的情感分析、情感本体构建、微博分析和大数据研究四个方面的相关工作展开。

第三章对微博结构化数据挖掘。对中文微博平台的网络结构、用户属性、文本特征等结构化数据进行研究。

第四章对本体结构和情感本体的构建进行研究，为微博情感分析奠定相应的基础。

第五章基于情感本体的微博特征识别和语义分析。

第六章基于情感本体的微博文本半结构化数据挖掘。在前面几章研究的基础上将基于语义和情感本体的情感分析技术应用于中文

试读结束：需要全本请在线购买： www.ertongbook.com

微博平台，对微博平台的产品（服务）评价和公众情绪进行研究。

第七章基于情感分析的微博非结构化数据开发，将微博情感分析应用到销量预测系统和航空质量评测系统。

第八章对本书的整体工作做了总结分析，指出其中存在的“瓶颈”和不足，并展望相关领域下一步可能的研究方向。

本书在完成过程中要特别感谢博士期间美国加利福尼亚州州立大学圣马科斯分校商学院何绍义教授和同济大学王洪伟教授两位导师提供的帮助及指导，感谢我的家人在日常生活中给予我的支持，感谢我现在的工作单位湖州师范学院，也要感谢卢小生编审对本书出版所做的细致的工作。

希望本书在理论上能丰富中文微博文本情感分析的研究体系，在实践中对基于微博的大数据挖掘和应用有一定的价值。

史 伟

2017 年 10 月

摘　要

随着计算机和网络技术的快速发展，互联网日渐成为各种信息的载体。人们在上面（包括新闻评论、产品评论、情感微博、网络社区等）主动地获取、发布、共享、传播各种观点性信息。这些观点性内容对于电子商务、舆情控制、信息检索等都具有重要的意义和实用价值。对网络文本的观点性内容进行自动情感分析成为 Web 信息处理的一个热点。而情感分析技术作为一个复杂的任务仍然面临巨大挑战，目前大多数研究采用统计自然语言处理技术，但这种方法对文本的语义理解能力不足，从而造成情感分类的准确率低，鲁棒性也不理想。同时对于当前最火热的中文微博文本的情感分析和大数据挖掘还处于初始状态。针对中文微博文本，探索从语义和情感本体的角度构建比较完整的大数据挖掘技术，旨在为中文领域的用户、企业、政府等相关方提供更为方便和科学的中文微博文本挖掘工具。

本书进行的主要创新性研究工作包括：

（1）本体结构及隶属度确定方法研究。本书将模糊描述逻辑应用在本体构建中，提出一个五元组的模糊本体模型，由概念集、角色集、实例集、模糊断言集和模糊关系集构成。基于 Tableaux 算法思想并结合算例，给出模糊本体模型推理问题的求解方法。针对模糊本体模型的隶属度计算问题，结合 Google 搜索，利用标准化谷歌距离（Normalized Google Distance，NGD）演算法，以实时且完全在线处理的方式计算关键词的相关性，并最终转化为模糊本体中的隶属度。

（2）基于知网的情感本体构建研究。本书针对微博文本情感表达的多样性和模糊性，将情感本体划分为评价词本体和情感词本体，利用模糊理论和知网相关概念，借鉴本体结构及模型，构建情感本体的基本模型。运用模糊化处理和语义相似度的相关理论，分别对评价词模糊本体和情感词模糊本体的情感类型和隶属度进行了相应处理。并通过与点互信息等方法比较，验证了情感本体模型在自动获取情感类方面的有效性，最终获得包括 6862 个评价词和 2090 个情感词的情感本体库，并进行了相关数据统计。

（3）微博文本的特征识别与语义分析。本书从语义的角度，运用构建的情感本体，对中文微博文本的特征识别与情感语义问题进行研究。介绍了情感空间模型，挑选出八种基本情感类和两种观点评价类进行标注。基于模糊情感本体，建立了文本中的产品（服务）特征提取方法，构建了情感类和强度标注规则，分析了程度词、否定词、连接词、修辞方法和标点等语义元素在情感分析中的影响。建立了从句子层到文档层的情感语义计算方法。最后通过相关语料进行实验，结果表明，建立的情感语义分析方法具有优良的准确性和应用性，同时分析了不同评论语料中情感具有的不同表达形式和关联关系。

（4）微博文本的半结构化数据挖掘。基于中文微博的产品评论分析，运用规范化的 TFIDF 加权方法提取产品特征，结合已建立的模糊评价词本体和 BMI 方法进行了产品特征评价词提取，建立了微博文本的影响力计算方法，结合模糊评价词本体和微博文本中的语义因素构建了微博中产品评论的情感类型和强度计算方法，最后通过实验和数据分析发现，本书构建的方法在各方面的表现都处于不错的水平，并具有很好的应用性。

基于中文微博的公众情感分析，以新浪微博为平台分析突发事件后公众的情绪状态和变化，抽取 2011 年 7 月 23 日“动车事故”发生后公众发表的微博并进行情感分析。提取八维情感类，应用已构建的模糊情感词本体，建立了微博文本的影响力和情感强度计算

方法，对“动车事故”后的公众情感随事态发展的变化进行探讨，分析发现，本书建立的方法具有很好的应用价值，可以为政府和相关部门通过微博进行舆情监测和分析提供参考。

（5）微博文本的非结构化数据开发。本书基于微博的情感分析进行了产品销量预测，通过构建的模糊情感本体和微博中的语义因素对微博文本进行情感计算，并将情感信息融入自回归模型中，建立自回归情感预测模型，根据以往的票房和观众在微博中的情感表现对电影将来的票房进行预测，通过实验与其他模型（未考虑情感因素）比较，发现本书的方法较其他方法具有更好的准确性和应用性。

关键词：情感分析　情感本体　大数据挖掘　微博文本　产品评论

ABSTRACT

With rapid development of computer and information technology, the internet is becoming the carrier of a variety of information. People obtain, publish, share and disseminate various opinioned informations on news commentary, product reviews, emotional microblogging, online communities, etc.. These opinioned contents have great significance and practical value for e – commerce, public opinion control, information retrieval, etc. Automatic sentiment analysis of viewpoint content of Web text has become a hot topic of Web information processing. Up to now, the sentiment analusis technique is still a complicated task with great challenge, the majority of studies use statistical natural language processing technology, but the semantics comprehension of the text is not enough for this method, resulting in low accuracy of emotion classification, robustness is not satisfactory. Meanwhile, sentiment analysis of hot Chinese microblogging is still in the blank state. This paper will focus on Chinese microblogging texts, explores to build a complete sentiment analysis technique based on semantic and emotion ontology, design to provide more convenient and scientific Chinese microblogging text mining tools for users, businesses, governments and other stakeholders in Chinese areas.

(1) Ontology structure and its membership determination. Ontology can formally describe concepts, terms and relationships of the special domain, but it can not express the fuzzy information. To solve the problem of the fuzzy information's description in ontology, fuzzy description logic was

employed in the construction of ontology. A 5 – tuple fuzzy ontology model was established, including roles set, examples set, fuzzy assertion set and fuzzy relationship set. Based on Tableaus algorithm idea and case studies, the paper gives the solution method of the reasoning problem of fuzzy ontology. The correlation between key words was processed with the new NGD algorithm by means of Google search, and convert the results into membership degree of fuzzy ontology at last.

(2) Study on construction of fuzzy emotion ontology based on How Net. The book builds fuzzy emotion ontology. For the diversity and ambiguity of emotional expression in Chinese microblogging texts, using fuzzy theory and ontology concept to build the basic model of fuzzy emotion ontology based on HowNet. Based on each features of the evaluation words and emotional words, using the relevant theories of fuzzy processing and semantic similarity, respectively processing the emotion classes and membership of the fuzzy ontology of evaluation words and emotional words, and by comparison experiment, our fuzzy emotion ontology outperform Pointwise Mutual Information in the aspect of getting emotion classes automatically.

(3) Feature recognition and semantic analysis of microblogging texts. From the semantic perspective, the book uses the constructed fuzzy emotion ontology, studies on sentiment analysis of Chinese microblogging reviews. At first, an emotional expression space model is described, and emotion semantic vocabulary is divided into emotional words and evaluation words, 8 emotional classes and 2 opinion evaluation classes are selected for manual annotation. Then, features of commodity (services) are extracted, emotion class and intensity are annotated based on the emotion fuzzy ontology, degree words, negative words, rhetoric, punctuation and other semantic elements are analyzed in texts. The method of sentiment calculation based on fuzzy semantic from the level of sentence to document is built. Finally, we do experiments on mobile phone reviews and wedding

试读结束：需要全本请在线购买： www.ertongbook.com

photography reviews, the results show that the established sentiment semantic analysis method has excellent accuracy and application, emotions have different forms of expression and association in different review corpus.

(4) Sentiment class and intensity analysis of microblogging texts. Analysis of product reviews based on microblogging, normalized TFIDF weighting scheme is applied to extract the most informative noun patterns to represent the product features, then we use the established fuzzy evaluation words ontology and BMI method to extract evaluation words for product features, and build the influence calculation method of microblogging text, construct the sentiment computing method of product reviews in microblogging, and finally we find that our system shows remarkable performance improvement over the baseline menthod through experiments and data.

Analysis of public sentiment based on microblogging, in this article, we perform a sentiment analysis based on all public microblogging posts about 'motor car accident' broadcasted by Sina microblogging users between July 23 and August 1, 2011. We extract eight dimensions of sentiment, build sentiment fuzzy ontology for sentiment analysis, establish sentiment intensity computing method of microblogging, explore the change of public sentiment after "motor car accident".

(5) The application of microblogging sentiment analysis – products sales forecasting. The book uses sentiment information in microblogging to forecast products sales, sentiment information in microblogging is analyzed through fuzzy emotion ontology and semantic factors in microblogging, and then sentiment information is incorporated into the autoregressive model (AR), the autoregressive sentiment – aware model is established. According to the box office in the past and sentiments in the microblogging, we predict the box office of the film in the future. We compare our system

with alternative models that do not take into account the sentiment information, as well as a model with a different feature selection method. Experiments confirm the effectiveness and superiority of the proposed approach.

Key Words: Sentiment Analysis Sentiment Ontology Big Data Mining Microblogging Texts Product Reviews

目 录

第一章 引言 …… 1

第一节 研究背景及意义 …… 1

一 互联网大数据的产生 …… 1

二 文本情感分析的应用 …… 2

三 微博文本中的大数据 …… 4

第二节 研究现状分析 …… 5

一 文本情感分析 …… 5

二 微博研究 …… 7

三 本体建模理论 …… 7

第三节 研究目的和内容 …… 8

一 研究目的和价值 …… 8

二 本书的主要研究工作 …… 10

第二章 文献综述 …… 15

第一节 基于语义的情感分析研究综述 …… 15

一 主客观文本分类 …… 18

二 基于语义文本情感极性分类研究 …… 20

三 情感强度分类研究 …… 28

第二节 情感本体构建研究综述 …… 33

一 情感类划分研究 …… 33

二 情感本体构建研究 …… 35

第三节　微博研究综述 …… 36
一　微博本身研究 …… 36
二　以微博为平台的情感分析研究 …… 37
本章小结 …… 39

第三章　微博文本结构化数据量化分析 …… 40

引　言 …… 40
第一节　微博定义与平台介绍 …… 40
一　微博定义 …… 40
二　微博平台 …… 42
第二节　微博与微博文本的特点 …… 45
一　微博的特点 …… 45
二　微博文本的特点 …… 46
三　微博文本中的特殊符号 …… 47
第三节　微博用户结构和内容分析 …… 49
一　微博用户结构 …… 49
二　微博平台上的内容分析 …… 51
第四节　微博文本获取与相关计算 …… 54
一　微博文本获取方法 …… 54
二　微博文本影响力计算 …… 56
三　微博话题影响力和热度计算 …… 58
本章小结 …… 58

第四章　情感本体模型的构建方法 …… 60

引　言 …… 60
第一节　本体结构及隶属度确定方法 …… 61
一　相关研究工作 …… 61
二　模糊描述逻辑的构造 …… 63
三　基于 FDL 的本体结构及其推理 …… 64

四 基于NGD的本体隶属度确定 …… 69
第二节 基于知网的情感本体构建 …… 71
一 情感本体构建基础问题 …… 71
二 情感本体结构设计 …… 74
三 基于模糊理论的评价词本体构建 …… 76
四 基于语义相似度的情感词本体构建 …… 78
第三节 数据统计 …… 82
本章小结 …… 84

第五章 基于情感本体的微博文本特征识别与语义分析 …… 85

引 言 …… 85
第一节 情感空间模型 …… 86
第二节 特征识别 …… 88
一 产品特征评价 …… 88
二 特征词提取方法 …… 90
三 语料特征词提取 …… 91
第三节 情感特征标注 …… 93
一 基本词性标注 …… 94
二 句子划分方法 …… 95
三 产品特征标注 …… 95
四 情感类标注 …… 96
第四节 程度词与否定词语义分析 …… 97
一 程度词语义分析 …… 97
二 否定词语义分析 …… 99
三 程度词与否定词不同组合语义分析 …… 100
第五节 几种影响因子语义分析 …… 103
一 标点符号语义分析 …… 103
二 连接词语义分析 …… 104
三 修辞句语义分析 …… 105

第六节　不同粒度层情感语义分析 …… 106
一　句子层情感语义计算 …… 107
二　段落层和文档层情感语义计算 …… 107
第七节　实验及数据分析 …… 108
一　实验流程设计 …… 108
二　程度词和否定词检测窗口分析 …… 109
三　特征识别和情感语义准确性分析 …… 110
四　情感类统计和关系分析 …… 111
本章小结 …… 115

第六章　基于情感本体的微博文本半结构化数据挖掘 …… 117

引　言 …… 117
第一节　基于情感本体的微博产品评论分析 …… 118
一　基于 TFIDF 产品特征提取 …… 119
二　基于 BMI 评价词提取 …… 120
三　微博文本影响力计算 …… 121
四　产品特征观点与情感类型和强度 …… 122
五　产品评论情感值计算 …… 123
第二节　基于情感本体的微博公众情感分析 …… 124
一　相关研究综述 …… 125
二　公众情感分析方法构建 …… 127
三　公众情感分析数据与文本清理 …… 127
四　情感本体构建与文本影响力计算 …… 130
五　微博文本情感类型和强度 …… 131
第三节　实验及数据分析 …… 133
一　微博产品评论实验分析 …… 133
二　微博公众情感实证分析 …… 138
本章小结 …… 142

试读结束：需要全本请在线购买：　www.ertongbook.com