中南财经政法大学
青年学术文库

大数据视角下的观点挖掘

Opinion Mining from the Perspective of Big Data

余传明 ○ 著

中国社会科学出版社

中南财经政法大学
青年学术文库

大数据视角下的观点挖掘

Opinion Mining from the Perspective of Big Data

余传明 ○ 著

中国社会科学出版社

图书在版编目（CIP）数据

大数据视角下的观点挖掘／余传明著．—北京：中国社会科学出版社，2018.9

（中南财经政法大学青年学术文库）

ISBN 978－7－5203－3092－3

Ⅰ.①大… Ⅱ.①余… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2018)第 200232 号

出 版 人 赵剑英
责任编辑 徐沐熙
特约编辑 孙 红
责任校对 王凤光
责任印制 戴 宽

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

印刷装订 北京君升印刷有限公司
版 次 2018 年 9 月第 1 版
印 次 2018 年 9 月第 1 次印刷

开 本 710×1000 1/16
印 张 18.5
插 页 2
字 数 232 千字
定 价 58.00 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

本书受中南财经政法大学出版基金资助

《中南财经政法大学青年学术文库》
编辑委员会

前　言

随着信息技术的快速发展，计算机互联网逐渐成为人们表达观点、情感的重要渠道。网络平台上的观点、评论等主观信息迅速增长，对这些信息进行分析能够帮助企业和公司改进产品与服务，及时修复可能潜在恶化的客户关系，提高企业在市场中的竞争力，因而具有非常重要的理论与实践意义。

在这种情况下，越来越多的企业和公司把关注投向互联网上的产品和服务评论，分析这些评论中所传递的重要信息。然而，由于评论信息的数量庞大且非结构化，通过人工阅读的方式往往难以完成，如何使用观点挖掘技术来解决海量的评论信息与个人有限的阅读能力之间的矛盾，已成为研究者亟待解决的重要问题。

为了尝试解决上述问题，我于2008年开始申报国家自然科学基金项目“WEB2.0环境下基于本体学习的观点挖掘研究”，并展开了相关的研究。本书最初构思于2008年秋，从提出想法到最终完稿，历时近十年。其间几度停笔，深感学术道路艰辛与不易。感谢家人、同事和朋友的鼓励和支持，使我能够克服重重困难，走出低谷，也使本书最终得以完成。

本书共分十二个章节。

第一章导论部分概述了大数据视角下观点挖掘的相关研究与不足，提出了大数据环境下所面临的规模跨度、领域跨度以及语言跨

度等挑战，从而引出了本书的研究问题。

第二章论述了大数据环境下观点挖掘的研究方法，揭示了本书的研究思路和研究框架，包括多领域多语言网络评论的下载、虚假评论的识别、产品名称和属性的识别、观点的极性判断、观点挖掘的领域适配、观点挖掘的语言适配、观点挖掘的规模适配、观点摘要、观点主题分析及可视化展示等。

第三章论述了虚假评论的识别问题。从评论利益相关者内容与行为特征相结合的角度出发，提出了一种基于个人、群体和商户的主体关系模型，包括虚假评论识别的行为指标体系、虚假评论者的主体关系建模、模型的参数确定、有效性评估以及模型的适应性分析等，并进行了相应的实证研究。

第四章论述了产品名称识别问题。以验证中文命名实体识别中的歧义消除问题为切入点，将产品名称识别转换为序列标注问题，对基于最大熵模型的产品名称识别和基于条件随机场模型的产品名称识别进行了探索，并进行了相应的实证研究。

第五章论述了产品属性识别问题。在分析现有产品属性识别方法不足的基础上，对两种无监督学习方法（即自组织映射方法和潜在狄利克雷分布模型）和一种监督学习方法（即支持向量机模型）进行了探索，并进行了相应的实证研究，对识别效果进行了检验。

第六章论述了观点的情感分析问题。使用统计和规则为基础的方法对旅游评论进行情感分析，提出了三组新的结合特征选择函数和传统的 TF－IDF 方法的方法，并通过扩充情感词和表情词等来补充原有的情感词典，在此基础上进行了多项实证研究。

第七章研究了观点挖掘的领域适配问题。以跨领域情感分析作为研究任务，提出一种跨领域深度循环神经网络模型，实现不同领域环境下的知识迁移，并与传统的栈式长短时记忆网络模型、双向长短时记忆网络模型、卷积神经网络—长短时记忆网络串联模型以

及卷积神经网络—长短时记忆网络并联模型进行了实证比较。

第八章研究了观点挖掘的语言适配问题。在不同语言文档使用不同词向量进行表示的基础上，提取与词项最相关的特征，并对这些特征重新编码，以此作为对跨语言环境下词向量所表达语义的一种补充，提出了跨语言的卷积神经网络模型，并进行相关的实证研究。

第九章讨论了观点挖掘的规模适配问题。论述了基于 Hadoop 平台的分布式计算框架 MapReduce 及 Spark 分布式计算，针对大数据集的情感分类问题，分别给出了在 Spark 分布式计算框架下决策树和逻辑斯蒂回归等算法的分布式实现，并进行了相关的实证研究。

第十章以自然语言处理技术与深度学习为切入点，探索了观点摘要的相关思路和研究问题，包括信息抽取方法、主题与语义分析方法、统计机器学习方法和深度学习方法等。

第十一章以西非埃博拉爆发的微博作为研究对象，探索了观点挖掘的主题分析的相关思路和研究问题，包括主题时序分析和演化等。

第十二章对全书进行了总结。

本书由我制订内容框架并审定全稿，同时负责全书大部分内容撰写与修改。陈百云、左宇恒、冯博琳参加了第三章关于虚假评论识别的研究，黄建秋老师、郭飞参加了第四章关于产品名称识别的研究，张小青、陈雷参加了第五章关于产品属性识别的研究，安璐教授、冯博琳参加了第六章和第八章关于情感分析以及观点挖掘的跨语言适配的研究，原赛、王峰参加了第九章关于观点挖掘的规模适配的研究，朱星宇、郑智梁参加了第十章关于观点摘要的研究，安璐教授、李纲教授、杜廷尧和周利琴参加了第十一章关于观点主题分析的研究。

本书系国家自然科学基金项目“大数据环境下基于领域知识获取与对齐的观点检索研究”（项目号：71373286）的研究成果之一。

余传明

2018 年 5 月于武汉

目　录