



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

余传明



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

版权所有 侵权必究

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

主 任：杨灿明

副主任：吴汉东 姚 莉

委 员：（按姓氏笔画排序）

朱延福 朱新蓉 向书坚 刘可风 刘后振

张志宏 张新国 陈立华 陈景良 庞凤喜

姜 威 赵 曼 胡开忠 胡贤鑫 徐双敏

阎 伟 葛翔宇 董邦俊

主 编：姚 莉

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 and 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 and 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

第一章 导论	(1)
第一节 观点挖掘:研究的兴起	(1)
第二节 从小数据到大数据:观点挖掘所面临的挑战	(4)
第三节 大数据环境下的规模跨度问题	(6)
一 潜在语义索引方法	(6)
二 佩奇排名方法	(7)
三 映射/规约架构	(8)
四 SQL 与 Hadoop 相结合的方法	(9)
第四节 大数据环境下的领域跨度问题	(10)
一 共同特征选择	(10)
二 目标领域文档选择	(11)
三 查询词扩充	(11)
四 迁移学习	(12)
第五节 大数据环境下的语言跨度问题	(13)
一 多语词典构建	(14)
二 语料库对齐	(15)
三 用户反馈和用户行为	(16)
四 领域知识库对齐	(17)
第六节 本章结语	(18)

第二章 大数据环境下的观点挖掘研究方法	(19)
第一节 观点挖掘的形式化定义与研究思路	(19)
第二节 多领域多语言网络评论的下载	(22)
第三节 评论的过滤与分类	(23)
第四节 产品名称和产品属性识别	(25)
一 关联规则法	(25)
二 点互信息法	(26)
三 概率潜在语义分析法	(27)
四 潜在狄利克雷分布法	(27)
五 相关主题模型法	(28)
六 最大熵原理法	(29)
第五节 观点极性判断	(30)
一 基于 WordNet 的方法	(31)
二 基于连接词的方法	(32)
三 基于点互信息的方法	(32)
四 松弛标记法	(33)
五 条件随机场法	(34)
第六节 领域跨度下的观点挖掘	(35)
第七节 语言跨度下的观点挖掘	(37)
第八节 规模跨度下的观点挖掘	(38)
第九节 观点摘要、主题分析与可视化展示	(41)
第十节 本章结语	(42)
第三章 虚假评论识别	(44)
第一节 虚假评论识别的意义	(44)
第二节 虚假评论识别的相关研究	(47)
第三节 虚假评论识别的行为指标体系	(51)
一 评论个人行为的指标体系	(52)

二	评论者群体行为的指标体系	(54)
三	商家行为的指标体系	(55)
第四节	虚假评论识别的主体关系建模	(56)
一	商户—个人(M-U)关系模型	(57)
二	个人—群体(U-G)关系模型	(58)
三	群体—商家(G-M)关系模型	(58)
四	虚假度迭代流程	(59)
第五节	虚假评论识别的实证研究	(60)
一	实验数据	(60)
二	参数确定及有效性评估	(60)
三	实验分析	(64)
四	与其他方法的对比分析	(66)
第六节	本章结语	(66)
第四章	产品名称识别	(68)
第一节	产品名称识别的问题描述	(68)
第二节	基于最大熵模型的产品名称识别	(73)
一	最大熵模型的理论基础	(73)
二	最大熵模型的参数估计算法	(74)
三	实验数据准备	(75)
四	最大熵模型的特征构建	(76)
五	最大熵模型的特征模板	(77)
六	特征生成	(80)
七	训练与测试	(80)
八	实验结果与分析	(81)
第三节	基于条件随机场模型的产品名称识别	(84)
一	利用条件随机场模型为产品名称识别问题建模	(84)
二	参数估计	(85)

三	模型求解	(86)
四	软件工具的选择	(87)
五	语料库构建	(87)
六	选取特征与特征模板	(88)
七	模型训练与测试	(88)
八	模板对产品名称识别效果的分析	(90)
九	语料库对产品名称识别效果的分析	(94)
十	与其他模型的识别效果比较	(96)
第四节	本章结语	(97)
第五章	产品属性识别	(98)
第一节	产品属性识别的问题描述	(98)
第二节	基于自组织映射的产品属性识别	(100)
一	自组织映射的原理	(100)
二	自定义的属性叠加矩阵及其原理	(101)
三	基于属性叠加矩阵的产品属性识别	(102)
四	网络数据收集	(103)
五	分词与词性标注	(104)
六	SOM 输入矩阵的构造	(104)
七	SOM 训练	(105)
八	SOM 的输出分析	(105)
第三节	基于 LDA 模型的产品属性识别	(109)
一	LDA 模型的原理	(111)
二	基于 LDA 模型的评论热点识别	(112)
三	数据预处理	(113)
四	输入向量的构造	(113)
五	模型求解	(114)
六	实验结果与分析	(115)

第四节 基于 SVM 模型的产品属性分类	(119)
一 支持向量机的原理	(119)
二 基于支持向量机的产品属性识别	(121)
三 网络数据收集	(122)
四 分词与词性标注	(122)
五 主观性标注与产品属性标注	(123)
六 输入矩阵的构建	(124)
七 模型的训练	(125)
八 实验结果及评价	(128)
第五节 本章结语	(129)
 第六章 观点的情感分析	(131)
第一节 观点极性分析的问题描述	(132)
一 特征选择及特征权重的研究	(132)
二 基于统计与基于规则的情感分类方法	(134)
第二节 基于改进的 TF-IDF 权重算法的情感分类	(136)
一 特征选择方法	(136)
二 数据集	(137)
三 评价标准	(137)
四 数据预处理	(138)
五 使用支持向量机的情感分类结果	(138)
第三节 基于情感词典和规则的情感分类	(143)
一 情感类别	(143)
二 情感辞典的构建	(143)
三 分类规则	(144)
四 使用规则组合的情感分类实验	(146)
第四节 本章结语	(152)

第七章 观点挖掘的领域适配	(154)
第一节 相关研究	(155)
一 跨领域情感分析	(156)
二 循环神经网络	(158)
第二节 研究问题与方法	(160)
一 研究问题的形式化定义	(160)
二 CD-DRNN 模型结构	(160)
三 对比方法	(163)
第三节 试验及分析	(168)
一 数据集	(168)
二 实验结果	(169)
三 讨论	(176)
第四节 本章结语	(177)
第八章 观点挖掘的语言适配	(179)
第一节 研究现状	(181)
一 基于机器翻译的方法	(181)
二 基于特征概率分布的方法	(182)
三 基于平行语料的方法	(183)
四 基于深度学习的方法	(183)
第二节 研究问题、模型与方法	(185)
一 研究问题及相关定义	(185)
二 先验特征的获取	(186)
三 模型结构	(188)
四 模型训练方式	(190)
第三节 实验及分析	(193)
一 数据集	(193)
二 比较方法	(193)

三 参数设置	(195)
四 实验结果	(197)
第四节 本章结语	(200)
 第九章 观点挖掘的规模适配	(201)
第一节 规模适配问题的提出	(201)
第二节 规模适配平台	(204)
一 Hadoop 平台	(204)
二 Spark 平台	(207)
第三节 规模适配算法	(209)
一 并行决策树算法	(209)
二 并行逻辑回归算法	(210)
三 并行朴素叶斯算法	(211)
四 并行随机森林算法	(212)
五 并行支持向量机算法	(214)
第四节 实验及分析	(215)
一 数据集与实验环境设置	(215)
二 评价指标	(216)
三 实验结果	(217)
四 讨论	(222)
第五节 本章结语	(223)
 第十章 观点摘要	(225)
第一节 信息抽取方法	(226)
一 图模型方法	(226)
二 篇章分析方法	(227)
三 结构模板方法	(228)
第二节 主题与语义分析方法	(229)

第三节	统计机器学习方法	(231)
第四节	深度学习用于观点摘要	(233)
一	序列到序列神经网络模型	(234)
二	注意力机制	(236)
三	先验知识	(237)
四	语义相关性	(238)
第五节	本章结语	(239)
第十一章	观点主题分析	(240)
第一节	研究问题	(240)
第二节	相关研究	(241)
一	微博主题分析	(241)
二	微博时序分析	(242)
三	微博可视化分析	(244)
第三节	主题演化模式和时序趋势的方法设计	(246)
第四节	实验过程与结果分析	(248)
一	数据描述和预处理	(248)
二	英文埃博拉微博的主题分析	(248)
三	中文埃博拉微博的主题分析	(253)
第五节	本章结语	(259)
第十二章	总结与展望	(260)
参考文献		(263)