信息系统协会中国分会第七届学术年会

（CNAIS 2017）

论 文 集

主办单位： 信息系统协会中国分会（CNAIS）

承办单位： 复旦大学 **管理学院**

支持单位： 国家自然科学基金委员会

清华大学出版社

中国·上海 2017 年 10 月 20-22 日

信息系统协会中国分会第七届学术年会

（CNAIS2017）

中国·上海 2017 年 10 月 20-22 日

大会荣誉主席：

陈国青（清华大学）

大会主席：

毛基业（中国人民大学）

程序委员会领域主席（以姓名拼音为序）：

黄丽华（复旦大学）
李一军（国家自然科学基金委员会）
马费成（武汉大学）
王刊良（中国人民大学）
徐 心（清华大学）
杨善林（合肥工业大学）

程序委员会（以姓名拼音为序）：

蔡 剑（□京大学）
陈福集（福州大学）
陈晓红（中南大学）
党延忠（大连理工大学）
葛世伦（江苏科技大学）
郭迅华（清华大学）
黄丽华（复旦大学）
黄 伟（西安交通大学）
霍佳震（同济大学）
寇 纲（西南财经大学）
李 纲（武汉大学）
李敏强（天津大学）
李一军（国家自然科学基金委员会）
梁昌勇（合肥工业大学）
刘 渊（浙江大学）
吕廷杰（□京邮电大学）
鲁耀斌（华中科技大学）
马费成（武汉大学）
毛基业（中国人民大学）
戚桂杰（山东大学）
邵培基（电子科技大学）
孙建军（南京大学）
王刊良（中国人民大学）
谢 康（中山大学）
徐升华（江西财经大学）
徐 心（清华大学）
严建援（南开大学）
杨善林（合肥工业大学）
叶 强（哈尔滨工业大学）
张朋柱（上海交通大学）
张 新（山东财经大学）
赵 晶（中国地质大学（武汉））
赵捧未（西安电子科技大学）
仲伟俊（东南大学）

程序委员会执行主席：王刊良（中国人民大学）

组委会主席：黄丽华（复旦大学）

组委会副主席：徐云杰（复旦大学）

院长/系主任论坛主席：左美云（中国人民大学）

主办单位：信息系统协会中国分会（CNAIS）

承办单位：复旦大学 **管理学院**

支持单位： 国家自然科学基金委员会　清华大学出版社

目录

ID 与注册号对应。可以通过本文件书签快速查找文章内容。

ID	论文标题	作者	页码
002	IT 外包与云计算情境下的 IT 业务匹配过程研究	邓春平 宋琦 王滢 孙玥璠	1
003	云制造平台工业企业准入指标体系研究	聂会星 邓旭	8
005	国内社科领域大数据研究知识扩散特征分析	曹玲 吴其静	15
007	移动虚拟社区治理对组织公民行为影响研究——从关系质量视角出发	迟铭 毕新华	22
008	农产品物流金融融通仓盈利模式演化博弈分析	杨波 吴燕	34
009	基于 S-O-R 模型的青年消费者“双十一”购物意愿形成机制研究	陈传红 李小倩	41
010	基于科学知识图谱的国内政策评估可视化研究	宁静 王亚民 马续补	48
011	比特币矿工的最优选择策略设计	郑阳 杜荣	53
014	基于贝叶斯网络推理的干散货疏运偷盗风险分析	崔维平 黄磊 宋容嘉	61
015	UGC 社媒前向商业变现动因研究	郭明君 吴俊 欧阳书凡	67
016	当出版遇上互联网+——中南传媒集团“互联网+”的实践	许媛 张美娜 余艳	74
017	基于依存句法关系的在线评论情感属性的降维	王洪伟 蔡文嘉	84
018	话题情感对众筹投资者行为的影响研究	疏卉 张昊 尤薇佳	93
019	电子商务与企业绩效：供应链敏捷与集成的中介作用	张馨 李琦 张宁	101
020	食品在线个性化定制情景下营养信息呈现方式对用户行为的影响研究	富露 周良 王刊良	105
021	风险驱动的智慧建筑系统开发案例——基于万达	张克迪 傅湘玲 齐佳音	117
022	iPad 和 Kindle 在电子阅读器市场的竞争与合作	姚忠 徐彪	123
023	基于论据的电子商务人机谈判模型	曹慕昆 庞俊杰	131
024	社会化推荐研究综述	王刚 蒋军 王含茹	137
025	基于知识元的科技文本资源组织方式研究	刘杰 秦春秀 赵捧未 刘怀亮	147
026	大数据背景下威胁评估对网络隐私顾虑的影响：组织隐私政策的调节	曲静 谢卫红 屈喜凤 张延林	154
027	在线评论对 O2O 模式创新扩散的影响研究	杜宾	164
028	技术压力对用户满意度的影响机制研究——角色压力的中介作用	王玮 喻亚琴 宋宝香	171
031	移动 O2O 模式下消费者购买意愿影响因素分析——基于公平理论和 BRA 模型的实证研究	刘百灵 孙文静 夏惠敏 徐伟	178
032	不确定需求下双寡头软件厂商两时期发布策略	王宇 李敏强 冯楠 陈富赞 田津	184

033	企业社交工作平台对工作—非工作边界影响研究	孙元 吴丽霞 潘绵臻	190
034	活跃用户影响力对投资组合绩效的影响——基于雪球网投资社群的探索	张偲媞 吴俊 殷雯	195
035	共享单车扩散过程中规范的采纳行为研究	马源鸿 曹云忠 刘佩雯 李敏	202
037	大数据与企业绩效关系研究——基于资源与分析洞察能力视角	谢卫红 钟苏梅 苏芳 王永健 王田绘	208
038	基于 CNKI 网络口碑领域文献的计量分析	张欢 张宁 尹乐民 李娜	219
039	一种考虑评论特征权重的在线评论子集提取方法	倪乃晨 张瑾 任明	226
040	基于公众兴趣的政务微博传播效果因素研究	冯小东 汤志伟 张会平	232
041	基于二部图的电子商务退货风险预测研究	张亮 刘冠男 马宝君	239
042	社交媒体用户倦怠与消极使用：基于扎根理论的探索性研究	李旭 张冰倩 刘鲁川	246
043	特定兴趣领域的社交媒体用户影响力研究	滕德宁 芦鹏宇	254
044	Optimal Dynamic Pricing of Online Platforms with Network Externalities	Guofang Nan, Lin Chen, Tianyu Wang, Mingqiang Li	260
048	双边网络外部性下平台企业最优产品线设计	金治州 冯海洋 李敏强	266
049	三维全景漫游系统用户使用行为影响因素研究	胡丽雪 方佳明 邵培基 曹云忠	272
050	降价式拍卖消费者学习效应研究——以贡天下为例	徐亚男 杨波 杜书彧	277
051	企业大数据治理模式的多案例研究	田金英 杨波 胡梦可	289
052	网络降价式限量拍卖中的一元效应研究——以贡天下为例	陈宇菲 杨波 张晶	297
053	降价式拍卖中消费者重复购买的影响因素研究	王星 杨波 马茜	306
054	网约车转型中的司机激励因素探讨	侯婷 程絮森	315
055	众筹项目融资绩效影响因素的计量经济分析——以淘宝众筹为例	张文涛 闫相斌 陈越	319
056	不同阶段微博口碑情感对票房影响的研究	俞一凡 黄京华 宋婷	326
057	金钱激励能导致更多的在线用户参与行为吗	匡丽妮 杨涵 颜志军	332
058	虚拟品牌社群人际互动对产品购买决策的影响	赵晓燕 沈波	337
059	基于隐私设计的信息系统安全研究	步飞 王能民	343
060	品牌转换意愿的形成机理研究：基于品牌和替代者的双重视角	丁晓燕 张新 张戈	349
061	产品众筹资助人与发起人情绪与融资动态关系研究	齐子豪 郑海超 李立婷 李万庚	355
062	基于 TOPSIS 的知识密集型众包任务人才选择模型实证研究	赵杨 袁析妮	362

063	支持数据驱动决策的情境建模：基于本体的方法研究	宋容嘉 王英 崔维平 黄磊	369
064	信息系统实施后工作特征及系统特征变化对员工工作绩效的影响机制研究	王玮 肖春华 宋宝香	377
066	基于组织承诺与组织控制的信息安全遵从研究	刘晨晖 王能民	384
067	基于 SERVPERF 模型的电商快递服务质量评价指标研究	金丹 任烜毅	394
068	在线问答社区信息价值影响因素的扎根理论分析	孙晓宁 齐云飞 赵宇翔 朱庆华	399
069	基于知识管理的电子政务信息资源建设	常金玲 刘青青	406
074	生产成本依赖规模经济效应的供应链需求扰动管理	李陶然 刘东苏	413
075	中国大陆内外信息资源管理研究现状分析	张甦 王宇	419
077	Understanding Interpersonal Work Connections and Personal Relationships between Co-workers though Electronic Communication Networks: A Case Study in China	WANG Youying, HUANG Qian, Robert M.DAVISON	425
078	虚拟社会资本对网红名人商业价值的影响分析研究	黄月涵 谷钰 洪帆	435
079	基于博弈论的企业信息安全技术评价	李晓彤 李华 杜黎	447
080	企业微信影响力及服务质量研究：以物流企业为例	杜松华 柯晓波 朱琳	452
081	基于 Word2vec 的微博多类别情感分析及电影票房预测	关琳 俞一凡 黄京华	458
082	协同消费参与意向的社会多维价值影响因素研究	蔡舜 彭志伟 张意成 庞晓 丁国维	464
083	基于深度学习 CNN 与协同过滤的在线商品推荐方法	管悦 陈国青 卫强	474
084	患者如何选择线上医生——医生贡献和患者回馈视角	洪紫映 张韦 邓朝华 刘汕	481
085	社会化商务意愿的实证研究：S-O-R 视角	甘春梅 林恬恬 许嘉仪	487
086	云计算对上市公司财务绩效影响的实证研究	李正华 王念新 葛世伦	493
087	基于语义查询扩展的关联主题推荐研究	霍辰辉 刘东苏	500
088	微信朋友圈用户潜水意向影响因素研究	刘晓丹 闵庆飞 刘子龙	506
089	智慧居家养老感知数据预处理研究	左美云 侯静波 蒋立新	512
090	基于机器学习的公益众筹项目融资能力评价模型研究	赵杨 武立强 谭道勋	521
091	电子政务云吸收与价值影响机制研究	梁乙凯 戚桂杰 周蕊	528
092	对未来关系的重视与 IT 企业创新合作	曾维君 艾宏峰	537
094	农产品电子商务顾客满意度和忠诚度的影响因素——基于产品类型的对比研究	李婷 罗志梅 林家宝 江飞	542
095	制度环境下企业农产品电子商务吸收的影响因素分析	李蕾 罗志梅 胡倩 林家宝	548

096	负面评论的商家回复与潜在消费者购买意向的关系研究	徐琬月 李瀛 陈昊 李文立	554
097	社会化媒体在多语言翻译资源构建中的应用研究	刘伟成 MISCHO William H.	561
098	我国“涉农”电子商务的研究回顾——基于 CSSCI 期刊的统计分析	李良强 邵培基 杨锐 曹云忠	566
100	移动信息技术与组织控制：文献综述与理论框架	周黎 杨琪 王祎	572
102	在线教育平台中竞争对手类型对用户学习绩效的影响研究	邓泓舒语 郭迅华 陈国青	583
103	社会化问答社区用户知识贡献的影响因素探究	刘子齐 吴鼎 郭迅华 李纪琛	590
104	基于深度学习的学业状态预测模型研究	王兴芬 孙彦超	597
105	外部压力对政府内部组织要素与开放政府数据质量的调节效应研究	赵玉攀 樊博	604
106	我国互联网境外上市公司的聚类、回归与演化分析	楼润平 孙鹏 毛彧	618
107	基于可变密钥强度的射频支付卡	姜琛凯 潘松洁	628
108	知识付费产品销量影响因素探究	石海荣 蔡舜 傅馨	641
109	XBRL 技术扩散与“一带一路”下财务报告的国际沟通研究	乔鹏程	651
110	基于使用行为分析的共享单车管理优化研究	傅哲 余力	662
112	社交媒体特性对于用户焦虑情绪影响的实验研究	张冰倩 李旭 刘鲁川	671
113	移动广告价值对消费者态度及购买意向的影响	黄丽娟 李林子 李成蹊	679
114	移动营销中消费者感知价值构成维度研究——概念界定与量表开发	黄丽娟 贾琳 李成蹊 李林子	690
115	信息交互能力：概念界定与研究框架	孙璐 李力 乐承毅	701
116	基于心理防御视角的个体网络知识分享研究	李玉豪 王刊良	715
118	传统企业的“互联网+”转型：公司治理与并购经营绩效	张悦悦	723
120	大体积耐用品适用于 B2C 电子商务吗？以家具为例	李英 陈振环 龚敏 赵越	731
121	网络疑病影响因素及对策研究	饶倩雯 李艳红	742
122	基于社会资本的知识基于社会资本的知识产品收入影响因素研究	邱婷 蔡舜 傅馨	747
123	基于微博内容分析的共享单车评价研究	张凌 罗曼曼 罗鹏程	753
124	基于 Citespace 的企业知识库研究知识图谱	李慧 田亚丹	758
125	智慧养老服务体系及平台构建研究	李彩宁 毕新华	766
126	产品众筹发起人性格对项目融资与执行结果的影响机理研究	李立婷 郑海超 刘琛	772
127	CIO 拼创行为对组织 IS 创新战略的影响研究	张延林 李礼 白海青 吴学雁	780
128	微信使用影响了生活满意度吗?基于社会资本视角的解释	郑大庆 黄林 王雨	786
129	国内基于本体的知识服务研究进展:核心内容	孙雨生 白璧娇 廖盼	791
130	共享房屋平台使用意愿的影响因素:基于 TAM 的实证研究	满溪柳 王洪伟	796

131	社交媒体中企业突发事件的舆情分析研究	周鹤 李良强 袁华 钱宇 侯伦	803
133	地理距离、投资经验与在线众筹投资决策:行业匹配度的调节作用	陈文波 宾颖 焦会玲	810
135	网络中立，补贴和内容创新	李枝勇 孙为政	817
136	权限请求界面设计:通过情境信息线索降低授权不确定性	柳君 刘子龙	826
137	社交媒体问题性使用行为的实证研究	王天华 刘子龙	833
139	基于在线农产品评论的消费者情感标签抽取方法研究	白梨霏 李开明 李良强 邹芳	840
140	公司披露文本分析研究进展	王洪伟 朱林源	847
141	在线短租房源图片对消费者行为意愿的影响	吴江 靳萌萌	853
142	基于网络数据的投资者情绪和股市相关性分析	刘赫 尚维 孙毅 汪寿阳	861
143	电子商务环境下多平台消费者价格偏好研究	张友莎 钱宇 袁华	867
144	全球数字贸易格局及影响因素研究:基于社会网络分析	陆菁 傅诺	872
146	职业认同与职业倦怠对教师网络实践社用户采纳意愿的影响-基于情感的中介作用	徐光 刘鲁川	882
147	在线客户评论的属性提取与细粒度情感分析:基于深度学习和层次聚类方法的视角	马宝君 陈璐 万岩	891
149	项目描述的欺诈性与众筹投资意愿：基于文本分析的方法	沈倪 王洪伟	899
152	基于旅游数字足迹的游客时空特征研究——以南京为例	廉同辉 余菜花	904
153	移动信息技术与组织结构、员工授权的跨层次关系研究	杨琪 周黎 王祎	911
154	基于在线招聘大数据的劳动力市场分析	刘耘	924
155	大学生对智能手机平面广告认知偏好研究	柴亚青 李慧	930
156	企业业务软件的云化迁移决策研究	何梦娇 任南 苗虹	937
157	网络效应在信息产品“免费 +增值”模式中的作用机制	李伟	945
159	基于贝叶斯网络的食品安全风险预警研究	莫名垚 姜同强	950
160	网站质量测量评估及其影响的元分析研究	叶许红 韩芳芳 陈慧栋	957
161	基于层次结构知识元的文本资源语义空间	秦春秀 李祯静 刘杰 谢庆球	963
162	制造企业 IT 能力对创新绩效的影响：数字化转型的中介作用	池毛毛 王伟军 陈秋阳	972
163	信息技术对科研论文引用的影响:基于经济学发表的实证	马鹏飞 张丹煜 张诚 徐云杰	982
164	颜值对网上交友影响的研究	张琦 张诚	988
165	新兴技术背景下的高校教育变革	赵钊 贺荣戈	995
166	基于云技术新型架构的育种数据服务平台	岳媛 赵刚	1000
167	基于 SNA 和 DMR 的慢病社群探测与主题演化趋势研究---以高血压为例	周利琴 潘建鹏 张斌	1006
168	众筹中顾客参与和产品市场表现的关系研究	苏颖 林丽慧	1016
169	产品共享对企业定价策略的影响研究	丁灵 廖貅武 杨橹	1023
170	社交媒体中食品安全风险沟通效果研究	史丰源 何德华 方雯琳	1030

171	SHARING DATA ON FITNESS APPLICATIONS: THE IMPACT OF SOCIAL FACTORS	ZHOU Ya, KANKANHALLI Atreyi, PHANG Chee Wei	1039
172	G-S 匹配机制在招聘网站的可用性研究	付自强 孙永洪 姜红丙	1045
173	远程问诊中不良医患关系的影响因素研究	晏梦灵 谭鸿瀛 梁嘉熠 李晨昱	1053
174	人民币国际化之区域经济联动与金融科技发展	王超 王坚	1061
175	全球政府开放数据在数字经济中的贡献评估报告	范佳佳	1069
176	基于 TAM 的众包物联网模型在灾难响应中的应用	黄虎 韩水华	1081
177	基于深度学习 LSTM 的电商销量预测研究	武玉英 严勇 何喜军 蒋国瑞	1088
178	物流信息系统“不当可视化”对顾客信息焦虑影响研究	李曼宁 袁月 董宜斐 于美婷	1097

在线客户评论的属性提取与细粒度情感分析：基于深度学习和层次聚类方法的视角*

马宝君[1]，陈璐[1]，万岩[1]
(1. 北京邮电大学 经济管理学院，北京 100876)

摘　要：随着电子商务的迅猛发展，网购已经开始成为人们生活中必不可少的一部分，电商平台上的在线客户评论内容对于消费者购买决策以及商家商品和服务改进都发挥着越来越重要的作用，如何能够自动、快速、有效地从大量的在线客户评论文本数据中提取相对完整的属性特征及进行细粒度情感分析，也成为电商平台信息服务提供商越来越关注的问题。鉴于此，本文提出了一种新颖的自动提取在线客户评论属性及其基础上的细粒度情感分析方法，通过应用句法分析模型提取候选属性词及其对应的语义关系，运用 word2vec 词向量模型训练得到语料中各个词所对应的词向量，并对候选特征词进行层次聚类得到商品属性特征维度，进而计算商品在各个属性维度上的情感强度。最后本研究通过天猫商城游戏本商品的实际数据实验分析验证了该方法的合理性与有效性。

关键词：属性提取；细粒度情感分析；层次聚类；深度学习；情感强度

1　引言

近年来，电子商务的迅猛发展带来了消费者行为模式的深刻改变，网购已经开始成为人们生活中必不可少的一部分。根据中国互联网信息中心的最新数据显示，截至 2017 年 6 月，我国网络购物用户规模已经达到了 5.14 亿，较 2016 年底增长 10.2%[1]。随着网络购物市场的日益繁荣，加之信息传播方式的巨大变革，在线客户评论的重要性愈发显著：对于消费者而言，在线客户评论是一项重要的信息来源[2]，对其购买决策过程有着重要的影响；对于在线商家而言，在线客户评论表达了用户对于所购买商品的评价与态度，是重要的反馈信息，能够为商家或商品的进一步完善提供指导作用。因此，消费者会借助在线客户评论来全面真实地了解商品的详细信息，从而有效降低网购过程中的购物风险；而商家则会针对在线评论进行积极响应以求保证消费者对于其商品或服务的好感，或者对于差评所产生的不良影响进行补救[3]。总体来说，无论是对于消费者还是商家，在线客户评论都发挥着重要作用。

目前主流的电子商务平台上积累了大量的在线客户评论，平台服务提供商也通过多种方式展示买家购买、体验商品后的态度、意见和评价，最常用、最有效的方式为展示评论中客户评价较多的商品属性及态度（如图 1 所示）。虽然上述展示方式较仅简单展示好、中、差评的数量、比例的方式对用户了解商品特点更方便有效，但也存在如下一些问题：（1）通过对若干评论数量不太多的商品在线客户评论的仔细梳理，可以发现电商平台上目前提供的评价属性及态度不太完整，即一方面评价属性不够全，另一方面涉及某一评价属性的评论条数不够完整，某些评论中的评价属性没有被提取出来（例如在淘宝中选中某一评价属性及态度会展示出涉及该属性及态度的“所有”评论）；（2）一般一个评价属性可能会涉及多种语义相关的表达或多个属性词，如果没有相关产品的领域知识，很难将较为完整的属性词都想到或提出来，这也会影响评论属性及态度展示的准确性和有效性，而相对准确的人工处理对有着大量在线客户评论的商品显然不具有可操作性。

(*a*)

(*b*)

图 1 京东(*a*)和淘宝(*b*)电商平台在线评论属性展示示例

*基金项目：国家自然科学基金(71402007, 71772017, 71471019).
作者简介：马宝君，男（汉族），副教授；陈璐，女（汉族），硕士研究生；万岩，女（汉族），教授。
通讯联系人：万岩，E-mail: wanyan@bupt.edu.cn

基于此，本文从深度学习和层次聚类方法的视角，提出了一种新颖的自动提取在线客户评论属性及其基础上的细粒度情感分析方法。通过运用句法分析模型从文本中提取得到候选属性词及其对应的语义关系，运用 word2vec 词向量模型训练得到语料中各个词所对应的词向量。随后，基于候选属性词、词向量以及情感词典，对候选特征词进行聚类，提取得到对应的各个属性特征维度；同时扩充情感词典，发现语料中的隐性情感词。最后，根据所得语义关系与情感词，计算商品在各个属性维度上的情感强度。同时，本研究选取了天猫商城 1099 种游戏本商品的在线评论数据进行实验，展示了初步分析结果并验证了该研究方法的有效性。

2 相关工作

由于本文所提出和运用的研究方法本质是运用层次聚类方法与 word2vec 模型，结合情感词典，对非结构化的在线客户评论文本数据进行自动的属性提取与细粒度情感分析，因此该部分的内容将围绕层次聚类方法、神经网络语言模型以及基于情感词典的情感分析三部分来展开论述。

2.1 层次聚类方法

层次聚类方法是将数据对象分为若干组并形成一个组的树形结构来进行聚类。层次聚类方法又可以分为自下而上和自上而下的层次聚类两种，而目前自下而上的聚合层次聚类方法比自上而下的分解层次聚类方法更加频繁地应用于实际的应用程序中[4]。自下而上的聚合层次聚类方法的思路是：初始将每一个对象作为一个聚类，然后将这些聚类按照一定的聚合条件聚合成较大的聚类，直到所有对象都聚合成一个聚类或者满足一定的聚合终止条件为止。聚合层次聚类方法的代表有 AGNES(Agglomerative NESting)方法[5]、BIRTH(Balanced Iterative Reducing and Clustering using Hierarchies)方法[6]、Chameleon 方法等[7]。

2.2 神经网络语言模型

统计语言模型的目标是学习语言中单词序列的联合概率函数，其最大的难点在于维度灾难。神经网络语言模型的出现则有效地解决了这个问题。最早的神经网络语言模型是由 Bengio 系统化提出的 NNLM（Neural Network Language Model）[8]，其基本思想是通过对模型的不断训练与优化，使得语料库中的每个单词获取一个分布式表示，不仅去除了维度灾难，还可以让模型能够了解语义相近的句子的数量。神经网络语言模型包含两大核心部分：一个是分布式表征[9]，即所谓的词向量（Word Embedding）;另一个是运用神经网络建立语言模型[10]。

神经网络模型在近十多年取得了较大的发展，所提出的各类模型也都是以 NNLM 作为模板，例如比 NNLM 更为简单的 CBOW 模型[11-12]、Skip-gram 模型[11-12]等。此外，为了使模型的训练更为快速有效，又出现了相应的 Hierarchical Softmax 算法[13]、Negative Sampling 算法[14]等。2013 年 Google 公司所开放的 Word2vec 深度学习工具则集合了上述模型与算法，能够快速有效地将语料库中词语以词向量的形式表示，获取词与词之间的语义相关性，为神经网络语言模型在各个领域的广泛应用提供了更为有效的方法[15]。基于 word2vec 的有效性，及其能准确捕捉词与词之间的潜在语义相似性，本文的研究方法即基于 word2vec 工具来计算文本的情感强度。

2.3 基于情感词典的情感分析

早在上世纪 90 年代初，国外就有相关学者通过情感词典进行文本的情感分析。Riloff 与 Shepherd 是最早进行相关研究的学者，提出了基于语料库数据来构建语义词典的方法[16]。Hatzivassiloglou 和 McKeown 在考虑了大规模语料数据中形容词语义情感倾向的限制性影响的基础上，尝试对单词的情感倾向进行判断[17]。在此之后，越来越多的研究开始关注情感词或情感短语与特征词之间的关联。Turney 等人则使用 PMI 方法，通过计算语料中非情感词的情感倾向使得情感词典得到了扩展，并运用语义极性算法（Semantic Polarity Algorithm）分析文本情感，最终取得了 74%的准确率[18]。

我国在这方面的研究也取得了较大进展。朱嫣岚等人基于 HowNet 情感词典，分别提出了基于语义相似度和基于语义相关场的词汇语义倾向性计算方法，判别准确率可达 80%以上[19]。李钝、曹付元等人从语言学的角度出发，采用“情感倾向定义”权重优先的计算方法获取短语中各词的语义倾向度，同时分析短语中歌词组合方式的特点，提出“中心词”的概念来对各词的倾向性进行计算来识别短语的倾向性和倾向强度[20]。白雪等人则在运用 word2vec 将语料中的所有词转换为词向量后，使用基于 PMI 方法改进的 SO-SD 算法计算词与情感词之间的语义距离，判断词的情感倾向，构建微博情感词典[21]。随后，基于所构建的微博情感词典，结合微博中表示程度的副词、感叹词、否定词以及表情图标等，对微博的情感倾向进行分类。其结果表明情感强度越高的微博，分类效果越好。

总体来说，基于情感词典的情感分析方法准确度较高，并且能够进行细粒度情感的相关研究。

然而局限于自然语言处理技术以及相关的数据提取方法，该类方法难以发现和获取数据当中的隐藏信息，其后续研究具有很大的发展空间。

3 研究方法与框架

本文所提出和应用的基于深度学习和层次聚类的在线客户评论的自动属性提取与细粒度情感分析方法包括如下六个模块（具体如图 2 所示）：

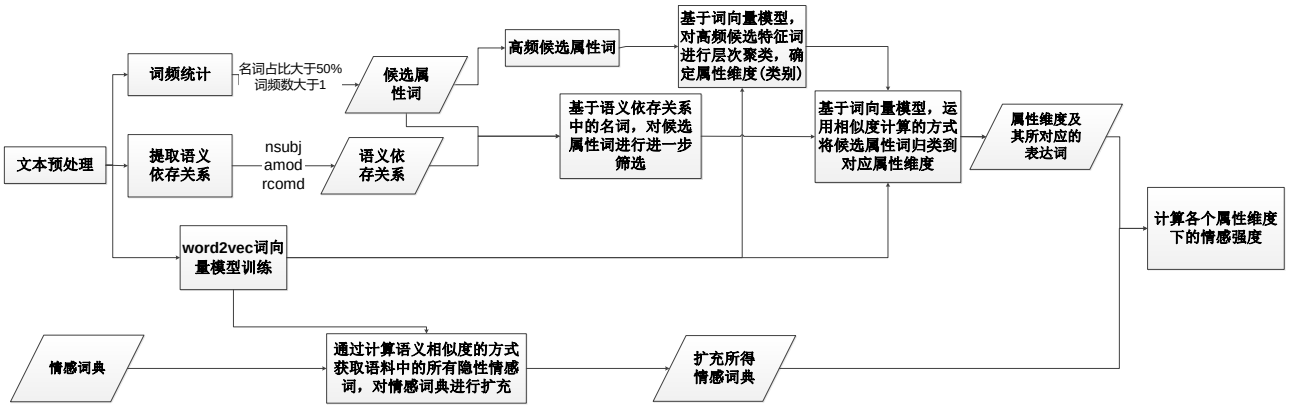

图 2　属性抽取和细粒度情感分析的研究框架及具体流程

（1）文本预处理与候选属性词筛选

对于获得的在线客户评论文本数据，首先需要进行必要的预处理，主要包含中文文本分词、词性标注以及去除停用词等操作。

在完成文本的预处理工作后，参照 Suleman[22]、Wu[23]以及 Hu[24]等人“属性词往往都是高频名词”的研究思路，对文本中的所有词语按照其所对应的词性进行词频统计，并从词频统计结果中选出名词标注占比大于 50%（因同一个词在不同的文本中可能有不同的词性标注结果）的所有词，将其作为候选属性词集合。

（2）语义依存关系提取

Manning 等人通过对各类语义依存关系的整理与研究，发现 nsubj(nominal subject)、amod(adjectival modifier)与 rcmod(relative clause modifier)三种语义关系常常包含了表达者对于某一事件或实体的观点或看法[25]。其中，nsubj 表示名词主语，amod 表示形容词表语，rcmod 表示关系从句修饰。Suleman 等人在研究中即通过找出这三种语义依存关系来实现对属性词的提取与筛选[22]。基于此，本研究从语料中提取这三种语义依存关系，并只保留两个词的词性分别为名词与形容词的语义关系。所保留的依存关系不仅将作为后续细粒度情感计算的重要基础，同时这些关系中的名词还可用来对候选属性词集合进行过滤，删除候选属性词集合中未出现在这些语义关系中的词语。

（3）词向量模型训练

作为一款词向量训练模型，word2vec 能够快速有效地将语料库中词语以词向量的形式表示，不仅可以消除词的歧义，同时还能准确获取词与词之间的语义相关性[26]。学者们曾在研究中证明了 word2vec 训练所得的词向量在语义分析上的有效性与优越性[26]。本研究即运用 word2vec 对语料进行词向量训练，获取各个词所对应的词向量。无论对于属性词聚类还是情感词典扩充，都需要基于词向量进行语义距离或语义相似度的计算。

（4）属性词归类

属性词归类包含两个步骤。首先，在完成候选属性词集合的过滤与筛选后，从中选出高频名词，并对这些词依据 word2vec 所训练得到的词向量，运用自底向上的层次聚类方法对其进行聚类得到若干属性类别。随后，计算候选属性词集合中其他属性词与各个属性类别之间的平均语义相似度，即该词的词向量与某一属性类别下各个词对应词向量的余弦相似度的均值，将其他属性词依次归类到对应的属性类别维度中。通过以上步骤，便可确定在线客户评论文本语料所对应的属性特征维度及其所对应的属性词列表。

（5）情感词典扩充

对于情感词典扩充的部分，可以从种子情感词典选择与隐性情感词计算两部分进行介绍。对于种子情感词典的选择，本研究选用了台湾大学情感词典(National Taiwan University Sentiment Dictionary, NTUSD)。由于大部分情感分析研究都是针对诸如推特、微博、在线评论等网络文本展开，而 NTUSD 中的词汇则来源于 General Inquirer

的中文翻译和《中文网络情绪词典》(Chinese Network Sentiment Dictionary, CNSD)[27]，更加适用于网络环境下的情感分析。

而在隐性情感词计算方面，本研究则运用了Bai 等人所提出的 SO-SD 算法[21]，具体处理策略介绍如下：结合种子情感词典，找出所有文本数据中所包含的情感词，即显性情感词。同时，将在线客户评论文本数据导入到 word2vec 模型中进行训练，获取文本中所有词的词向量形式。遍历除去显性情感词以外的其他所有词，通过计算余弦相似度的方式，找出与该词最为相关的若干个词，并观察这些词中是否包含显性情感词，如果包含则认为该词可能在表达上具有一定的情感，并作为隐性情感词保留，反之则认为该词是中性词。随后，对于所有隐性情感词，运用文献[21]中所提出的 *SO-SD* 方法判断隐性情感词的情感倾向，并将所求得的值作为对应隐性情感词的情感强度。*SO-SD* 的计算公式如下：

$$SO-SD(word) = \sum_{pword\in Pwords} SD(word,pword) - \sum_{nword\in Nwords} SD(word,nword) \quad (1)$$

其中 *pword* 表示与 *word* 最为相关的若干个词中所包含的某个正向显性情感词；*Pwords* 表示与 *word* 最为相关的若干个词中所包含的全部正向显性情感词；*nword* 表示与 *word* 最为相关的若干个词中所包含的某个负向显性情感词；*Nwords* 表示与 *word* 最为相关的若干个词中所包含的全部负向显性情感词。

此外，其中：

$$SD(word_1,word_2)=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^{\ 2}}\sqrt{\sum_{k=1}^{n}x_{1k}^{\ 2}}} \quad (2)$$

式（2）中 n 为词向量的维度，x_{1k} 为第一个词向量中第 k 维度的值，x_{2k} 为第二个词向量中第 k 维度的值。

对于计算得到的 *SO-SD* 值，我们采用 p 和 q 作为判断阈值：

$$\text{如果 } SO-SD(word)\begin{cases} >q & \text{则}word\text{为正向隐性情感词} \\ \in[p,q] & \text{则}word\text{为中性词} \\ <p & \text{则}word\text{为负向隐性情感词} \end{cases} \quad (3)$$

（6）属性情感值计算

对于各商品在线评论中所对应的所有语义依存关系，从中提取得出对应的名词与形容词，根据名词所属的属性类别以及形容词在情感词典中所对应的情感倾向，分别统计该商品下各个属性类别的正向情感词数量与负向情感词数量，并将两者相减的值作为该商品在该属性上的细粒度情感值。

4 数据实验

4.1 实验数据采集与预处理

为了展示该研究方法的有效性，我们选取了2016 年 11 月 1 日前后天猫商城上所有的 1099 种游戏本商品及针对它们的所有在线客户评论文本数据作为研究对象，来探究商品特征属性以及其上的细粒度情感分析。由于天猫商城平台在线评论是通过异步加载的方式显现出来的，因此本研究运用基于 AJAX 的定址网络爬虫技术抓取了 2016 年 11 月 1 日前后天猫商城中所有游戏本商品的对应在线评论文本数据，并选取了评论日期在 2016 年 10 月 31 日之前的所有数据，共计 137,570 条，包括 117,051 条初始评论及 20,519 条追加评论，对应商品 1099 件。

随后，对于所抓取的文本数据分别进行分词、词性标注以及去除停用词等预处理工作，将每一条评论文本表示成由若干个词语的向量形式，且每个词都包含其对应的词性。在此过程中，由于在线评论均是以中文形式表达，因此使用中科院的 ICTCLAS 工具包进行分词处理，使用斯坦福大学研发的 Stanford Postagger 对分词后的句子进行词性标注，使用哈工大的停用词词表以及词性标注去除句子中的停用词。

4.2 实验设置及结果分析

在候选属性词集合获取的工作中，针对我们已抓取的 1099 件游戏本商品的 137,570 条在线客户评论，基于词频统计与词性标注的方法，我们得到的候选属性词集合中共包含了 983 个词。用所提取语义依存关系中的对应名词对候选属性词集进行过滤，删除所有未出现在这些语义关系中的名词，则得到有效属性词 497 个。从这 497 个有效属性词中提取出名词对应词频大于 1000 的所有词共计 105 个高频属性词。

针对词向量模型训练，本研究选用了基于 CBOW 模型的 word2vec 深度学习工具。具体地，我们将词向量维度设置为 50，选用 Hierarchical Softmax 算法作为词向量优化方式，获取语料中每个词所对应的词向量。通过计算词向量的余弦距离，可以有效对词与词之间的语义相似性进行评估。限于篇幅，此处以语料中的高频名词“物流”、“价格”以及网络流行词汇“给力”、“呵呵”为例，分别获取与这些词语义距离最为相近的前 5 个词，所得结果如表 1 所示。从表 1 中可以发现，运用 word2vec 训练所得的词向量能够从本研

究所用语料文本中有效获取词与词之间的语义相关性。以“物流”为例，与该词语义相似度最高的前 5 个词均与其表达了相同或相似含义，如“快递”、“发货”等；而“给力”和“呵呵”则分别与“棒”、“大赞”等正向情感词以及 “骗人”、“倒霉”等负向情感词具有高度语义相似性。

表 1　使用 word2vec 得到的语义最相似词示例

高频词	语义最相似词	语义相似度
物流	快递	0.8074778
	物流速度	0.7946379
	发货速度	0.7714251
	发货	0.7583793
	顺丰快递	0.7005256
价格	价钱	0.6567913
	价位	0.64149135
	经济	0.58467937
	便宜	0.5749178
	性价比	0.5716945
给力	快	0.5959953
	棒	0.55695117
	大赞	0.54730815
	很快	0.53670025
	神速	0.5325734
呵呵	醉	0.6997956
	倒霉	0.69149774
	呵呵呵呵	0.667705
	骗	0.6596455
	骗人	0.6530406

针对上述得到的 105 个高频属性词，以它们使用 word2vec 训练得到的词向量为基础，以词向量之间的余弦距离为词与词之间的语义相似度，运用自底向上的层次聚类方法对其进行聚类。在实验中，我们采用以聚类算法中常用的类内对象相似度之和除以类间对象相似度之和为聚类效果评估测度[28]，10 折交叉验证的方法来寻找最优聚类类别数，在此数据集上得到最优类别数为 16，高频属性词的聚类结果如表 2 所示。

表 2　高频属性词聚类结果

簇	属性词
簇 1	本本 本子 笔记本 机子 机器 游戏本 整体 电脑 品牌 产品 宝贝
簇 2	颜色 外形 外观 颜值
簇 3	商家 卖家 老板 店家 客服服务 客服态度 客服态度 服务 服务态度
簇 4	电池 小时
簇 5	正品 质量 实体店
簇 6	价位 价格 价钱
簇 7	速度 反应 运行速度
簇 8	win10 系统 win7 软件
簇 9	内存 硬盘 显卡
簇 10	分辨率 屏幕
簇 11	声音 风扇 温度 散热 散热器
簇 12	性能 配置
簇 13	鼠标垫 电脑包 鼠标 赠品 键盘
簇 14	发货 快递 物流 包装
簇 15	压力 lol 玩 lol 特效 游戏 玩游戏
簇 16	办公 视频 电影

根据对各个簇及其对应词语的分析，从游戏本商品的在线客户评论所表达语义内容的角度出发，可以认为买家主要从以下 16 个方面对游戏本商品进行评估：产品整体，外观，商家服务，电池续航，产品质量，价格，运行速度，系统与软件，硬件，屏幕，声音与散热，性能，赠品与配件，物流，游戏体验以及其他用途。在确定了游戏本商品的各个特征属性后，对于属性词集合中的其他属性词，依次计算其与各个簇中所包含词的平均相似度，并基于计算结果将该词划分到相似度最大的属性类别中。限于篇幅，我们在表 3 中仅展示了以上 16 个属性特别类别中的 4 个及其对应的属性词。

表 3　商品属性特征及对应属性词

属性特征类别	属性词
产品整体（簇 1）	电脑 笔记本 机子 宝贝 本本 机器 品牌 整体 本子 产品 游戏本 款电脑 牌子 笔记本电脑 台电脑 款笔记本 款机子 整机 款机器 一台电脑 台笔记本 款产品
商家服务（簇 3）	客服 卖家 服务 店家 服务态度 态度 客服态度 老板 商家 客服服务 店家服务 售后服务 客服服务态度 店主 指导 帮助 掌柜 幻影 诚信 雨舞 信息 客服人员 科长 录音机 技术人员 天虎 阳光妹子 旋风 月光 人员 悍将 工程师 店家态度 中途 个客服 花荣 工作人员 销售 客服幻影 技术员 mm 姐姐 哥哥 说客 贝贝 骑士 服务人员 答复 客服联系
性能（簇 12）	配置 性能 方面 电脑非常 电脑性能 电脑配置 电脑确实 待机 机子性能
游戏体验（簇 15）	游戏 玩游戏 lol 压力 特效 玩 lol fps 大型游戏 网游 魔兽 gta5 世界 ps 单机 飞车 个游戏 cf dnf 3d 单机游戏 剑灵 火线 撸啊撸 最高画质 主流游戏 网络游戏 dota2 古墓 刺客 游戏卡 坦克世界 逆战 英雄 款游戏 传说 信条 电脑玩游戏 玩 cf 任务 剑三 巫师 团战 wow 天涯 战地 大型单机游戏 玩 dnf 玩单机 ol dota fps100

从表 3 中可以发现，在“商家服务”与“游戏体验”属性特征类别所对应的属性词中都包含了许多看似与该类别不太相关的词语，如“花荣”、“贝贝”、“骑士”、”lol”、”dota”、”cnf”等。通过对相关在线评论文本的观察与分析，可以发现“商家服务”属性类别下的某些属性词，是对应了商家客服人员的昵称，如评论“特地先用几天才来评价的，机子颜值高，性能好，没什么大问题，有些小问题联系热心的客服花荣一切搞定，联想的网上售后还是不错的”，就表达了对客服人员花荣的好评；而“游戏体验”属性类别下的一些属性词则表示某款游戏的具体名称，如评论“电脑很不错，LOL 全高效稳定 60，使命 10，极品 18 都能玩，附上图”，就表达了用户使用该产品玩 lol 游戏的体验非常好。此外，对于一些量词与名词或名词与副词的结合，如“款电脑”、“电脑非常”等，是由于分词不准确所导致的，并不会影响到属性词的归类与细粒度情感的计算。

情感词典扩充的主要工作即是在语料中找出可能带有情感色彩但是并未出现在情感词典中的隐性情感词，并将该类词语添加到种子情感词典当中，作为针对该语料进行情感分析的重要基础资源。本研究选用 NTUSD 作为种子情感词典，运用 SO-SD 算法获取隐性情感词。

本实验中进行隐性情感词计算的思路如下：语料中的所有词进行遍历，判断该词是否包含在种子情感词典中。如果是，则跳过到下一个词；如果不是，则获取与该词语义相似度最高前 25 个词，判断这 25 个词中是否包含显性情感词。如果不包含，则认为该词为中性词；如果包含，则运用 SO-SD 算法计算得出该词的情感倾向值，并根据情感倾向值的大小以及所设定阈值区间来判断该词是否可以作为隐性情感词。此处所设定的[*p*, *q*]阈值为[-0.1, 0.1]，即如果计算所得情感倾向值在该区间内，则认为该词不可以被认为是隐性情感词。

按照上述处理策略，我们共从天猫游戏本在线客户评论文本语料中提取出了隐性情感词 11,602 个，其中包括 4,014 个正向隐性情感词以及 7,588 个负向隐性情感词。表 4 列出了情感倾向绝对值最高的各 10 个正负向隐性情感词及其对应的 SO-SD 情感倾向值。将所提取的所有隐性情感词加入到种子情感词典 NTUSD 中，即完成了情感词典的构建工作。在此基础上，针对商品的每个属性特征类别，分别统计各个属性类别的正向情感词数量与负向情感词数量，并将两者相减的值作为该商品在该属性特征类别上的细粒度情感值。

表 4　隐性情感词及其对应 SO-SD 值

正向 隐性情感词	SO-SD 值	负向 隐性情感词	SO-SD 值
经典	6.76430	一声不吭	-6.58744
优异	6.61272	马德	-6.4349
秀气	5.81988	大爷	-5.7168
更上一层楼	5.64711	闪屏	-5.21904
全新正品	5.50649	关门	-4.9266
满分满分	5.40054	关门大吉	-4.88805
梦寐以求	5.27354	闹心	-4.81394
诚恳	4.85831	裁	-4.81095
绚丽	4.81830	骂	-4.74916
正点	4.77006	慌	-4.69062

考虑到同一类型下的不同商品在各个方面往往会存在较为显著的差异；因此，消费者在购买商品时都会根据其自身偏好的不同，选择购买不同的商品。为了证明本实验所选取的游戏本商品属性特征类别之间存在显著的差异性，同时从侧面对本方法的合理性进行论证，本研究分别找出在所挖掘的 16 个属性特征上消费者满意度最高及最低的游戏本商品及其对应品牌（如表 5 所示）。

根据表 5 的结果，可以根据不同消费者的个人偏好来进行更有针对性的商品推荐。如对于更为注重游戏本商品外观及材质的消费者而言，可以推荐其购买 HP/惠普 PAVILION 15-bc011TX；对于更重视商品整体状况的消费者而言，可以推荐其购买 Asus/华硕 K K552WE；而对于渴望更快收到商品的消费者来说，则可以推荐其购买 Asus/华硕 X555YI 7110-554LXFA2X10 等。

5　结语

本研究基于层次聚类方法和深度学习的 word2vec 模型，提出了一种新颖的自动提取在线客户评论属性及其基础上进行商品属性细粒度情感分析的研究框架，并以天猫商城中所有游戏本商品为实验对象，结合其实际在线客户评论数据进行数据实验分析，验证了本文所提出研究方法的合理性与有效性。

基于所存在的一些问题，未来的研究可以针对以下两方面展开：（1）采用相关方法更为合理地标注和设置情感词的情感强度，使得情感强度的计算更为精确；（2）考虑到在线客户评论的短文本特性，后续研究可以考虑运用或设计更适合于短文本主题建模的相关模型，从而使得研究结果更为严谨客观。

表 5　不同属性特征的情感倾向值及其对应商品型号

属性特征	满意度（情感倾向值）最小值	对应商品型号	满意度（情感倾向值）最大值	对应商品型号
散热/声音	-17	炫龙 炎魔 T1 Pro	66	MACHENIKE T57 D6
性能配置	-2	HIPAA/海鲛 V5s-R2 Basic	128	炫龙 炎魔 T1 Pro
配件/赠品	-6	RABOOK/镭波 Firebat F760S1	84	炫龙 炎魔 T1 青春版
物流	-8	Dell/戴尔 灵越 15(7560) Ins15-1745G	395	Asus/华硕 X555YI 7110-554LXFA2X10
游戏体验	-4	神舟战神 T6-S5D1	177	炫龙 毁灭者 DC
其他用途	-5	HP/惠普 PAVILION 15-bc011TX	60	Asus/华硕 X555YI 7110-554LXFA2X10
整体	-6	Dell/戴尔 Vostro14-5480-5528	343	Asus/华硕 K K552WE
外观/材质	-15	Lenovo/联想 小新 310-14ISK	183	HP/惠普 PAVILION 15-bc011TX
卖家服务	-14	Lenovo/联想 小新 310-14ISK	178	Asus/华硕 X555YI 7110-554LXFA2X10
电池续航	-10	Asus/华硕 A a456UR7200	8	QRTECH 麦本本 大麦 3S
质量品质	-9	Dell/戴尔 Vostro14-5480-5528	151	Lenovo/联想 G510AM -IFI
价格	-2	Lenovo/联想 小新 310-14ISK	262	HP/惠普 PAVILION 15-bc011TX
运行速度	-8	ThinkPad E450 20DC-A07KCD	201	Asus/华硕 顽石 —FL5900U7500
系统	-18	Dell/戴尔 Vostro 14VR-1528	45	Lenovo/联想 小新 700-15isk I5
硬件	-14	Lenovo/联想 小新 310-14ISK	13	MACHENIKE T57 D1
屏幕	-20	Lenovo/联想 y50 Y50p-70-IFI	32	炫龙 炎魔 T1 Pro

参 考 文 献

[1] 中国互联网信息中心(CNNIC).第 40 次中国互联网络发展状况统计报告[EB\OL]. http://cnnic.cn/gywm/xwzx/rdxw/201708/t20170804_69449.htm, 2017-08-04.

[2] Chatterjee P. Online Reviews: Do Consumers Use Them?[J]. Advances in Consumer Research, 2001, 28.

[3] Ma B J, Yuan H, Wu Y. Exploring performance of clustering methods on document sentiment analysis[J]. Journal of Information Science, 2017, 43(1): 54-74.

[4] Han J W, Pei J, KamberJian M. Data mining: concepts and techniques (3rd ed.)[M]. Morgan Kaufmann Publishers, 2011.

[5] Kaufman L, Rousseeuw P J. Finding Groups in Data: an Introduction to Cluster Analysis[M]. John Wiley & Sons, 1990.

[6] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases[C]. Proceedings of ACM SIGMOD Conference, Montreal, Canada, pp. 103–114, 1996.

[7] Karypis G, Han E H, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling[J]. COMPUTER, 32:68-75, 1999.

[8] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Lan-guage Model[J]. Journal of Machine Learning Research, 2003, 3(6):1137--1155.

[9] Hinton G E. Learning Distributed Representations of Concepts[C]. In Proceedings of CogSci. 1986.

[10] Xu W, Rudnicky A I. Can artificial neural networks learn language models?[J]. In International Conference on Statistical Language Processing, 2000.

[11] Mikolov, Kopeck J, Burget L, et al. Neural network based language models for highly inflective languages[C]. Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Computer Society, 2009:4725 - 4728.

[12] Hinton G E, Mcclelland J L, Rumelhart D E. Distributed representations[J]. Parallel Distributed Processing Eds Rumelhart Et Al, 1986:77 - 109.

[13] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[J]. Aistats, 2005.

[14] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Eprint Arxiv, 2013.

[15] word2vector: https://code.google.com/p/word2vec/.

[16] Rilo E, Shepherd J. A corpusbased approach for building semantic lexicons[C]. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97), 1997:117--124.

[17] Hatzivassiloglou V, Mckeown K. Predicting the semantic orientation of adjectives. ACL[J]. Proceedings of the Acl, 1997:174--181.

[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] 朱嫣岚, 闵锦, 周雅倩,等. 基于 HowNet 的词汇语义倾向计算[C]. 全国第八届计算语言学联合学术会议（JSCL-2005）论文集. 2005:14-20.

[20] 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(4):132-134.

[21] Bai X, Chen F, Zhan S. A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec[C]. 2014 IEEE International

Congress on Big Data (BigData Congress), . IEEE, 2014:358 - 363.
[22] Suleman K, Vechtomova O. Discovering aspects of online consumer reviews[J]. Journal of Information Science, 2016, 42(4): 492-506.
[23] Wu Y, Zhang Q, Huang X, et al. Phrase dependency parsing for opinion mining[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3, 2009, pp. 1533-1541.
[24] Hu M, Liu B. Mining and summarizing customer reviews[C]. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.
[25] De Marneffe M, Maccartney B, Manning C D. Generating Typed Dependency Parses from Phrase Structure Parses[C]. Language Resources and Evaluation, 2006.
[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems, 2013, pp. 3111-3119.
[27] Ku L W, Liang Y T, Chen H H. Tagging heterogeneous evaluation corpora for opinionated tasks[C]. Language Resources and Evaluation, 2006.
[28] Zhao Y, Karypis G.. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, 2005, 10(2): 141-168.

Feature Extraction and Fine-grained Sentiment Analysis for Online Customer Reviews: from Perspectives of Deep Learning and Hierarchical Clustering Method

MA Baojun[1], CHEN Lu [1], WAN Yan [1]
(1. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract： With the rapid development of e-commerce, online shopping has become an indispensable part of people's life. The contents of online customer reviews on e-commerce platforms are playing an increasingly important role for both consumer purchase decisions and sellers' improvements of goods and services. Hence, how to extract relatively completed features automatically, efficiently and effectively from a large amount of online customer reviews and conduct fine-grained sentiment analysis, has attracted increasingly concerns of information service providers. In view of this, this paper proposes a novel approach to automatically extract features for online customer reviews and conduct fine-grained sentiment analysis, in which through the application of syntactic analysis model to extract candidate feature words and their corresponding semantic relationship, we utilize word vector model named *word2vec* to train and obtain the word vector for each word in the corpus, and conduct hierarchical clustering on candidate feature words to get the commodity attribute dimensions, and then calculate the emotional intensity for all dimensions of the commodity attributes. Finally, this paper conducts comprehensive real data experiments on real online customer reviews of all the gaming laptops on Tmall.com to verify the rationality and effectiveness for our proposed method.

Key words： Feature Extraction; Fine-grained Sentiment Analysis; Hierarchical Clustering; Deep Learning; Emotional Intensity